A COMPARISON OF AGGREGATION METHODS OF SUBJECTIVE PROBABILITY DISTRIBUTIONS

BY

YUHONG GU

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Arts in Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2009

Urbana, Illinois

Advisers:

Professor David V. Budescu, Fordham University
Assistant Professor Ali E. Abbas

Abstract

One of the goals of psychological research on subjective judgments is to develop procedures that can improve judgment aggregation quality. The need to aggregate various judgments arises since in many cases the decision maker is uncertain about the possible outcomes of his decisions solicits suggestions from multiple advisors.

In this paper, we study the quality of aggregation of multiple subjective probability distributions of future temperatures, using data collected by Abbas, Budescu, Yu and Haggerty (2008), as a function of 4 factors – the elicitation method (Fixed Probability versus Fixed Variable), the aggregation method (combining directly points on the distribution or aggregating parameters of fitted distributions), the aggregation statistic (using the mean or the, more robust, median to represent the aggregated values), and group size (we used data from 32 judges and we compare results of 200 replications of sub-groups of increasing size: the 32 single judges (n=1), 16 pairs of judges (n=2), 8 groups of n=4 judges, 4 groups of n=8 judges, 2 groups of n=16 judges, and a summary of all n=32 judges).

The quality of aggregation is measured primarily by the closeness of the estimated probability distribution to the reference distribution based on historical data. We observed that as group sizes increases, aggregation quality improves (closer fit to the historical values) and it matters less which judges are aggregated and how the judgments are aggregated. Aggregates based on FP assessment generate higher quality than aggregates based on FV assessment under most circumstances. When FP is adopted, point aggregation generates better results than parameter aggregation. If FV has to be adopted for practical reasons, using parameter aggregation with mean may produce higher quality results.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Award No. SES 06-20008 (Ali E. Abbas and David V. Budescu, Principal Investigators).

I wish to express my gratitude to my advisors, Dr. David Budescu and Dr. Ali Abbas, for the tremendous help and guidance they have offered to me during my master's study. Their professionalism and research dedication have been great examples for me to learn from. In particular, I would like to thank Dr. Budescu for continuing guiding me for an extended period to complete my master's thesis. The time and effort he spent editing and commenting on my thesis in very details, rounds after rounds, are sincerely appreciated.

I also appreciate the guidance from other faculty members in the Quantitative division, including my academic advisor, Sungjin Hong, as well as the others. The help and care I have received from my lab colleagues have helped me to grow both professionally and personally. Thanks to Steve Broomell for sharing his program with me, and to Andrej Dietrich for always patiently helping me with my statistics questions. Thanks to Carol Nickerson and Lori Hendricks for their help. Thanks to the department of Psychology for awarding me the fellowship during my first year of study at University of Illinois, and thanks to my advisors for offering me opportunities to work as a research assistant.

Table of Contents

CHAPTER 1 INTRODUCTION

On many occasions, decision makers (DMs) seek advisors when they are uncertain about the possible outcomes of their decisions and/or their corresponding likelihoods. The need to aggregate the various advisors' opinions arises since in most cases DMs solicit suggestions from multiple advisors.

Previous research has shown several benefits of aggregation. Aggregate judgments utilize more information – the effect of increasing the number of information sources is similar to increasing the sample size in an experiment. Aggregating judgments also reduces the impact of extreme estimates that may have resulted from faulty or inaccurate information. DMs aggregate judgments also because this is a more inclusive and ecologically representative process that can boost DMs' confidence in the credibility and validity of the final aggregates (Budescu, 2006). In one heavily studied context – probability judgments – there is an extensive literature (e.g., Harvey, Bolger & McClelland, 1994) that shows that in general, with just a few exceptions, subjective probability estimates are too extreme, implying overconfidence on the part of the judges. Ariely et al. (2000) pointed out that aggregating forecasts could provide a solution to this challenge and obtain more realistic and useful estimates.

It is useful to distinguish between the behavioral and the mathematical approach to aggregation of judgments. There are two types of behavioral aggregation approaches: one attempts to generate agreement among a group of experts by having them interact in some way, while under the other approach, a single DM contacts several experts independently and then combines their estimates (Fiedler, 1996; Yaniv & Kleinberger, 2000). In contrast, mathematical aggregation methods consist of processes or analytical models that operate on the individual probability distributions to produce a single aggregated probability distribution (Clemen and Winkler, 1999; Clemen and Winkler, 2007). Reviews of the literature on mathematical combination of probability include Cooke (1991) and French and Insua (2000).

Mathematical aggregation methods range from simple summary measures, such as arithmetic or geometric means of probabilities, to procedures based on axiomatic approaches or on various models of the information-aggregation process.

Early work on mathematical aggregation of probabilities focused on axiom-based on aggregation formulas. The strategy is to postulate certain properties that the aggregated distribution should follow and then derive the functional form of the aggregated distribution. The linear opinion pool (Stone, 1961; Abbas, 2009) where the aggregated probability estimate is the weighted average of individual judges' estimates is a typical example of this approach.

The Bayesian approach can be used to aggregate point probabilistic forecasts or probability distributions by using Bayes theorem to update the prior according to the various judges' estimates. However, the Bayesian approach is difficult to apply due to the complexity involved in the assessment of the likelihood function. In addition to accounting for the precision and bias of information provided by each judge, the dependencies between judges must be captured as well. Several works have been compared Bayesian and non-Bayesian approaches, including recent work by Budescu and Yu (2006).

The elicitation and expression of opinions in some quantitative form is essential for this process. The form in which the group's opinions are expressed influences, to some extent, the selection of a pooling method, since it would be natural to express the consensus judgment in the same form as the original forecasts (Genest and Zidek, 1986). The degree of dependence between judgments (Winkler, 1981; Wallsten and Diederich, 2001; Johnson, Budescu and Wallsten, 2001) and the number of estimates being aggregated (Ariely et al., 2000) are also important factors affecting the level of accuracy and diagnosticity improvement achieved by aggregation. Most recently, partition dependence, a judgmental bias that arises from the particular way in which a state space is partitioned for the purposes of making probability judgments, has been brought to researcher's attention. Clemen and Ulu (2008) developed a model which can reduce partition dependence. Bordley (2009) presented a new approach for combing probability assessments from different experts, which allows experts to assess their probability assessments across different partitions.

This work focuses on mathematical approaches, to aggregate the subjective probability distributions of continuous variables. We chose to examine the processes for combining whole distribution as it provides deeper insight than just mean-level analyses do. Instead of asking judges to give mean estimates, they estimated probabilities of several points on the distribution. Working with distributions of response times Rouder, Lu, Speciman, Sun and Jiang (2005) showed that the variance, shape and other information provided by distribution level analyses are useful for identifying the best fitting model. Engelberg, Manski, Williams (2007) also pointed out that point predictions such as mean or median tend to give a more favorable view than the means/medians/modes from subjective probability distributions.

The present study

We study the quality of various methods of aggregation multiple subjective probability distributions. The quality of aggregation is measured by the closeness of the estimated probability distribution to a reference distribution based on historical data. We study the effect of four factors on the quality of forecast aggregation: the distribution elicitation method, the aggregation method, the aggregation statistic, and the group size. Section 2 explains how the data were collected, discusses the four factors and the measures of aggregation quality analyzed. In Section 3 we present the analyses of the impact of the four factors on the quality of the aggregation. The last section summarizes the results regarding the quality of the various aggregation approaches.

CHAPTER 2 METHDOLOGY AND DATA

Data Collection

In early December 2006 Abbas, Budescu, Yu and Haggerty (2008) conducted an online probability elicitation study with 103 students (the vast majority of whom were enrolled in the Decision Analysis class at Stanford University). Sixty-four of these students reported their subjective probability distributions regarding the high temperature in Palo Alto a week before a target date (December 12th). In the present paper we analyze data of a subset of this sample.

Half of the students (randomly selected) were provided with a historical chart of the temperature. Previous analyses (Abbas, et al., 2008) have shown that the historical information did not affect any aspect of the results. Therefore, in the following discussion we disregard this factor and only differentiate between the methods by which the judges provided their estimates. Two distribution elicitation methods were implemented through a series of simple binary choices and all judges used both methods (presented in random order). Under the fixed probability approach (FP) they estimated the temperature values corresponding to 5 (cumulative) probabilities (5%, 25%, 50%, 75% and 95%). Under the Fixed Variable (FV) method they judged the probabilities corresponding to 5 temperature values. The temperatures spread symmetrically around the midpoint of the range selected by each judge. Hence the temperatures provided by the various judges were not the same, but for some analyses we normalize them to allow for meaningful comparisons.

Thirty-two of the sixty-four judges who estimated similar lower and upper bounds of the temperatures on that date, and whose judgments were relatively monotonic were selected for the analysis in this paper. Their subjective probability estimates will be aggregated in various ways that will be described below. Summary of key features of the assessments' of the selected judges (highest temperature, lowest temperature, ranges of temperatures, as well as Kendal correlations which measure monotonicity) are presented in Table 1.

	Mean	**Std Dev**	**Minimum**	**Maximum**
Kendal Correlation for FP	0.95	0.07	0.74	1.00
Kendal Correlation for FV	0.96	0.09	0.60	1.00

Range of temperatures	18.83	4.46	10.56	31.11
Lowest temperature	4.54	3.09	-1.00	8.89
Highest temperature	23.37	3.62	16.67	31.11

Table 1: Summary of the ranges of forecasts and the monotonicity of the forecasts of the 32 selected judges

We also collected high temperatures in Palo Alto for the week of December 12th (three days before and three days after) for 53 years (http://www.wunderground.com/history/airport/KPAO/2007/12/15/DailyHistory.html). These historical data are used to construct the reference distribution against which the estimates obtained from our judges are compared.

Methodology

Four factors

We study the impact of four factors on the quality of the aggregated forecasts.

1. Distribution elicitation methods: fixed probability or fixed variables

Spetzler and von Holstein (1975) identified three basic types of encoding methods: fixed probability (FP), fixed variable value (FV), and a mixture of the two. We use FP and FV data in our study. The FP method uses a fixed setting on the probability wheel (hence, FP) and asks judges for the value of the variable (the temperature in our study) whose cumulative probability corresponds to the wheel setting. In a typical application of the FP approach, one selects several cumulative probabilities (p) and judges are asked to report values (X) such that $F(X) = Pr(x \leq X) = p$. For example, in our study judges were asked "what is the temperature (in degrees) that you believe there is only a 75% chance that it will not be exceeded in Palo Alto on December 12th?"

The most commonly chosen quantiles of the cumulative distribution are the median (p = 0.5), and the quartiles (p = 0.25 and 0.75) (see for example Hora, Hora, & Dodd, 1992; Lichtenstein, Fichhoff & Phillips, 1982). Variants of this approach are widely used in practice when experts provide their High, Base, and Low values (0.1, 0.5, 0.9) for a variable of interest to construct decision trees, or tornado diagrams (see for example McNamee and Celona, 2001), or in assessing dose-response curves for various pollutants (see for example, Wallsten, Forsyth & Budescu, 1983). In some cases, analysts assess 5 quantiles (0.10, 0.25, 0.50, 0.75, and 0.90) or

even 7 quantiles (0.01, 0.10, 0.25, 0.50, 0.75, 0.90, and 0.99) (Lau, Lau, & Zhang, 1996). In this study 5 quantiles (0.10, 0.25, 0.50, 0.75, and 0.90) were used.

The second approach assesses the fractiles using a fixed value of the variable (hence, FV) and asks for the probability wheel setting that corresponds to the cumulative probability of that variable value. In practice, judges are asked to assess the probabilities Pr(x≤X) for several selected values of the target variables (x). In this study, judges were asked, for example, "What is the probability that the temperature of Palo Alto on December 12th is no higher than X=25 degrees?"

Abbas et al. (2008) found that the two methods were practically indistinguishable in some ways (e.g., the goodness of fit of the Beta distributions based on the FP and FV judgments). There was a slight, but consistent, superiority for the FV method along several dimensions such as monotonicity, robustness, and precision of the estimated fractiles. Also, judges were able to make FV judgments faster and were more likely to reach full indifferences (rather than establishing narrow intervals) for the fractile assessments, which suggest that the FV method is easier and more natural.

2. Aggregation methods: Points (non-parametric) or Parameter (parametric)

Rouder et al. (2005) compared two methods to generate group level distributions of response times: Vincentizing and Parameter averaging (PA). Vincentizing is a popular nonparametric method for constructing group level distributions (Heathcote, Popiel, & Mewhort, 1991; Vincent, 1912): Essentially, the individuals' estimates for each quantile are aggregated across the group. For example, consider two judges who provided estimates, X_1 and X_2 respectively, for the first quartile, i.e., $F(X_1) = F(X_2) = 0.25$. The aggregated estimate for the first quartile is some function of X_1 and X_2 (for example, their average).

Under the parameter aggregation method each judge's estimates are fit with a theoretical distribution and the distribution parameters are aggregated across the group. In our study beta distributions were fitted. For example, assume that judge 1's estimates were fitted

with $Beta(\alpha_1, \beta_1)$ and judge 2's estimates were fitted with $Beta(\alpha_2, \beta_2)$. The aggregated distribution will be some function of the two, for example $Beta(avg(\alpha_1, \alpha_2), avg(\beta_1, \beta_2))$.

Rouder et al. (2005) found that PA methods outperformed the Vincentizing based methods in terms of accuracy, measured by the root mean square error (RMSE) of the estimation errors (defined as the difference between the true value and the estimates of the aggregated parameters). According to Rouder et al., the reason for the superior performance of PA method is that for three parameter distributions, such as Weibull distribution (studied in their work), Vincentized estimates do not necessarily become more accurate with sufficiently large sample sizes (Rouder and Speckman, 2004; Thomas and Ross, 1980). However, PA methods can provide consistent estimates. This lack of consistency in estimates from Vincentizing explains its relatively weaker performance with larger sample sizes.

3. Aggregation Method: Mean or Median

For both methods, we combined the judgments by taking their average or their median. Research has shown that a simple average often provides better results than more complicated and sophisticated combination models (Guerard & Clemen, 1989). The advantages of using simple averages (the simple mean of the forecasts being aggregated) are that they are simple to apply, understand, and explain to the end-users of the forecast, regardless of their technical backgrounds.

One concern regarding simple averages is their sensitivity to extreme values. Subjective forecasts can sometimes vary considerably, due to several reasons including misunderstanding or mis-interpretation of the information, variance in the judges' backgrounds and so on. Researchers have considered the median as a less sensitive alternative option. However, the median has provided mixed results with regard to the aggregation quality. The performance of the median is better than the mean in Agnew (1985), worse in Stock and Watson (2004), and about the same in McNees (1992).

To find robust aggregation options, Jose and Winkler (2008) considered trimmed means and Winsorized means. These measures involve taking the i smallest and i largest forecasts and either

deleting them (trimming) or setting them equal to the (i+1)th smallest and (i+1)th largest forecasts (Winsorizing), respectively. More trimming or Winsorizing means that less information is being used (more data points are being deleted or adjusted), but the mean is less likely to be overly influenced by an extreme value. The mean and the median represent the extreme cases of trimming and Winsorizing: At one limit we have the mean, which corresponds to (0) no trimming or Winsorizing; at the other limit is the median, which represents (50%) maximal amount of trimming or Winsorizing. Jose and Winkler (2008) found that moderate trimming of 10–30% or Winsorizing of 15–45% of the estimates can provide improved aggregated estimates, with more trimming or Winsorizing being recommended when there is more variability among the individual estimates. When the level of trimming and winsorizing goes beyond the suggested percentage, performance gets worse. Their study also suggested that trimming and Winsorizing tend to be more robust than simple average as they reduce the risk of high errors measure by the percentage of times the symmetric APE exceeds certain values. Their results showed that mean outperforms the median in terms of accuracy, but is less robust.

4. Group size: number of subjective distributions being aggregated.

The accuracy of aggregation increases with every advisor added to the pool, but at a diminishing rate that depends on the inter-judge correlation (e.g., Ariely et al., 2000; Johnson, Budescu, & Wallsten, 2001; Wallsten & Diederich, 2001, Jose & Winkler, 2008). Many of these (and other) studies show that substantial improvement in predictions can be obtained with as few as 2 to 6 judges and that the rate of improvement above this numbers is reduced. The rate of improvement is a function of the inter-judge correlation (see analysis by Broomell & Budescu, 2009). This is particularly relevant in cases where the inter-judge correlations are high, as in our study. Figure1 shows the distribution of the (32 * 31 / 2 =) 496 Kendal and Pearson correlations between the 32 judges. They are all positive and very high.

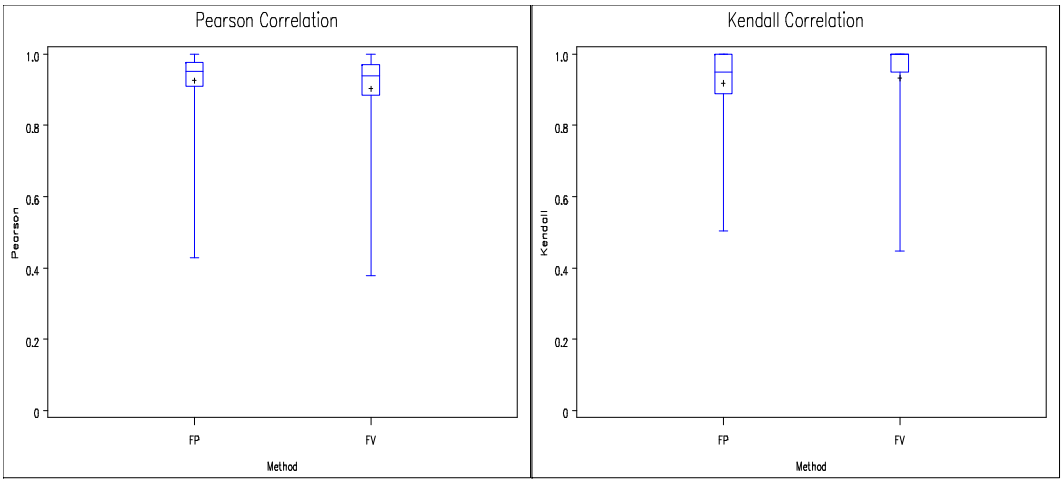


Figure 1: Distribution of the Inter-judge correlations for the FP and FV judgments

In our study, the size of group was doubled successively as judges were clustered into groups of size 2, 4, 8, 16 and 32. For each group size, 200 different ways of grouping were simulated to allow estimation of the variance of process. To implement this process we selected randomly 200 (out of the total 32! possible) permutations of the 32 judges. For each permutation selected we created groups of size 2 by combining the first and second judge (they are group 1), the third and fourth judge (they are group 2), and so on. To obtain groups of size 4 we aggregated the two previous groups of size 2, etc. The same process was applied to all other group sizes. Essentially, we generated 200*(32/2) = 3,200 groups of size 2, 200*(32/4) =1600 groups of size 4, 200*(32/8) =800 groups of size 8, 200*(32/16) =400 groups of size 16, and one group of size 32. Thus, all levels and types of aggregation use exactly the same amount of information.

Quality measures

The quality of aggregation was measured in several ways. First we show that the variance of the aggregated distributions decreases as the number of individual distributions being aggregated increases. Then, we evaluate the accuracy of the aggregated distributions by comparing them to the reference distribution based on the historical data.

We used two approaches: The first approach operates directly on data points without fitting distributions of any particular form. We evaluate the closeness between historical data and

estimated data using the Kolmogorov-Smirnov (KS) measure. The second approach operates on fitted distributions. Instead of paying special attention to the 5 points, we look at measures of the global quality of the aggregates by simply comparing the parameters of the various distributions with those of the reference distribution.

1. Kolmogorov-Smirnov Statistic (KS)

The Kolmogorov–Smirnov test (K–S) is a form of minimum distance estimation used as a nonparametric test of fit of probability distributions to a reference probability distribution. The KS statistic for a given cumulative distribution function $F(x)$ is the supremum of the absolute difference between the empirical distribution and the reference

$$KS = \sup_x |F(x) - F_{ref}(x)|$$

Smaller values of KS indicate that the empirical distribution is closer to the reference.

2. Measures of global quality of aggregates

We fit beta distribution to each judge's individual data, as well as grouped data. The beta distribution is a family of continuous probability distributions defined on the interval [0, 1] parameterized by two positive shape parameters, typically denoted by α and β. The expected value and the standard deviation of beta distribution are:

$$EV = \frac{\alpha}{\alpha + \beta}, \qquad SD = \sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}}$$

In this paper, we compare the EV and SD of beta distributions fitted to individual or grouped judgments, with the EV and SD of the reference distribution.

CHAPTER 3 RESULTS AND ANALYSIS

We fit beta distributions to both historical and empirical data. As shown in Figure 2, the beta distribution provided a good fit to the historical data. The mean and variance of the historical data are 14.45 (C) and 8.74 (C).

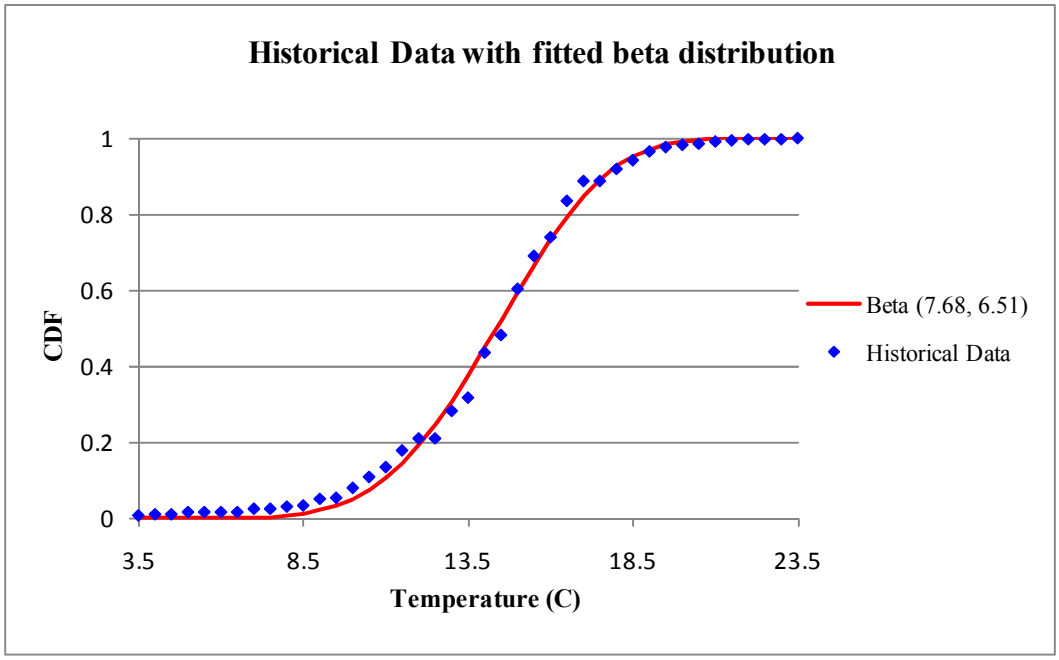


Figure 2: Historical data fitted with beta distribution: Beta (7.68, 6.51)

This distribution served as the reference distribution in our aggregation quality measurement. Below we explain how we aggregated subjective probability distributions.

For each judge, we have five data points $(X_{ij,}F(X_{ij}))$, where $i=1,2,3,4,5$ refers to the 5 points on the probability curve, and $j=1$ to 32 refers to the judge.

For the FP analysis, we have

$$F_1 = F(X_{1j}) = 0.05, F_2 = F(X_{2j}) = 0.25, F_3 = F(X_{3j}) = 0.5,$$

$$F_4 = F(X_{4j}) = 0.75, F_5 = F(X_{5j}) = 0.95,$$

When we group judges (for example if we use averaging), we have

$$X_{irn} = mean(X_{i,r,1}, X_{i,r,2}, \ldots, X_{i,r,n}),$$

n = group size = 1, 2, 4, 8, 16, 32, r = group ID = 1, 2, …32/n,

We then pair (X_{irn}, F_i) to build aggregated distributions. For the purpose of calculating quality measures, we also derived the corresponding CDF value $F_{ref}(X_{irn})$, on reference distribution for each X_{irn}.

For the FV analysis, we approximated the normalized X, denoted by $N(X)$ with certain values.

$$N(X_{ij}) = \frac{X_{ij} - X_{\min,j}}{X_{\max,j} - X_{\min,j}}$$

$$N(X_1) = N(X_{1j}) \approx 0.025, N(X_2) = N(X_{2j}) \approx 0.25, N(X_3) = N(X_{3j}) \approx 0.5,$$

$$N(X_4) = N(X_{4j}) \approx 0.75, N(X_5) = N(X_{5j}) \approx 0.875$$

After grouping, $F(X)_{irn} = mean(F(X_{i,r,1}), F(X_{i,r,2}), ..., F(X_{i,r,n}))$. We then pair $(X_i, F(X)_{irn})$ to build an aggregated distribution. For the purpose of calculating quality measures, we also derived the corresponding CDF values, $F_{ref}(X_i)$, on reference distribution for each X_i.

An example of two judges and a group formed by aggregating their individual judgments is provided in the appendix.

Fixed Probability Analysis

The effects of aggregation on the range of the distributions

In Figure 3, we display the maximal and minimal CDF value for each aggregated distribution, i.e., $\max\_F(X_{irn})$, and $\min\_F(X_{im})$ for each i. For example, for group size = 2, there are 16 groups * 200 replications = 3,200 groups that generate 3,200 aggregated distributions. The maximal and minimal CDF values for each of the 5 quantiles were selected from these 3,200 aggregates. We used linear interpolation between adjacent points to approximate the two bounding distributions. Figure 3 shows that the range of the aggregated distributions decreases monotonically as the number of distributions being aggregated increases, regardless of whether

we used the mean or median, and point or parameter aggregation methods. In other words, the more judges we aggregate, the less it matters which judges we combine.

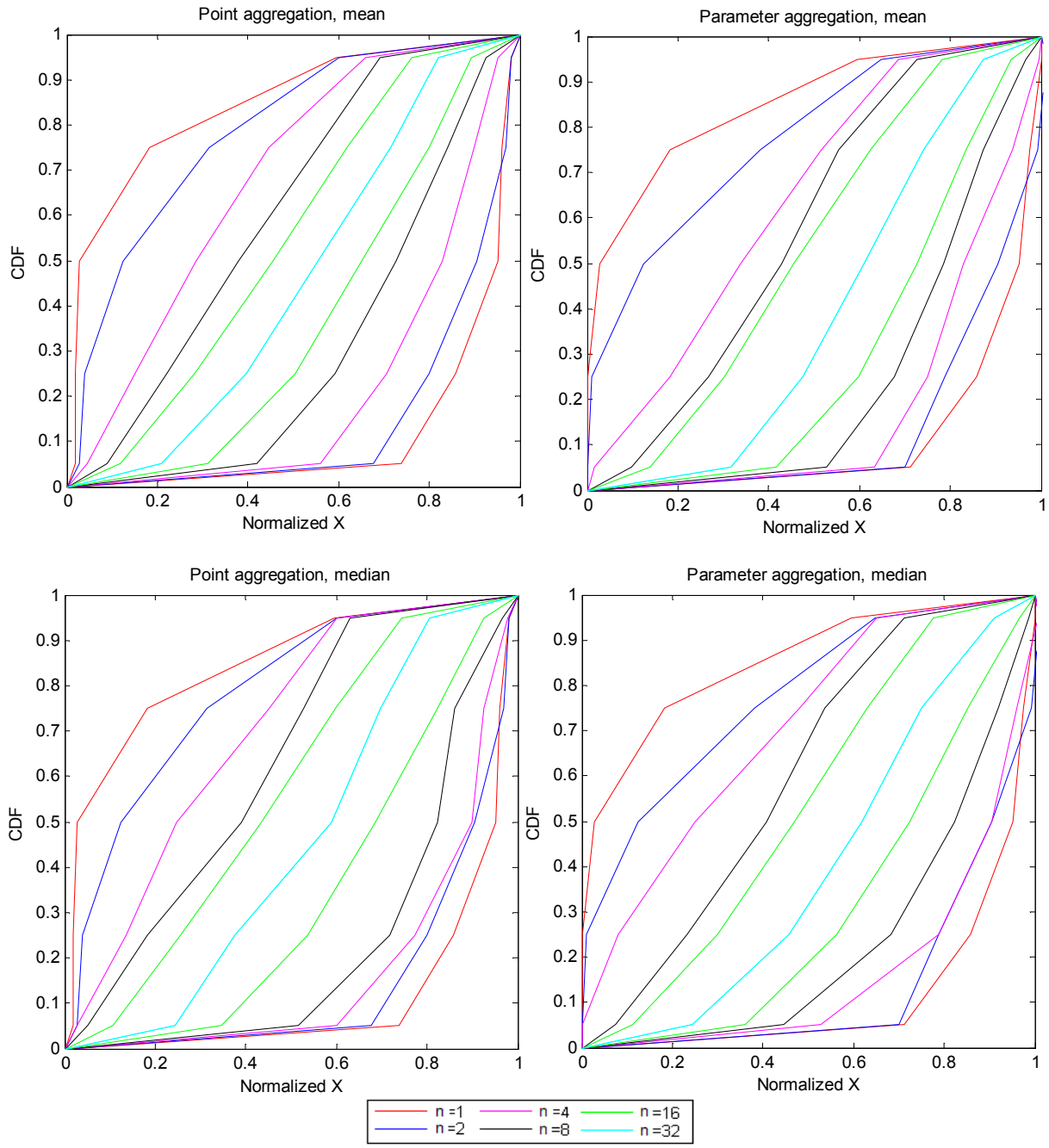


Figure 3: Bounding CDF distributions constructed from maximum and minimum CDF values of aggregated distributions as a function of group size for the FP assessments

Similarity to the reference distribution – Kolmogorov-Smirnov analyses

Box plots of the KS measures for all four combinations of aggregation method and statistics as a function of $\log_2$(groups size) are presented in Figures 4. Every step along the abscissa corresponds to doubling the group size. Increasing group size has two clear effects: (1) Reduction

in the mean and median of the distributions, i.e. closer fit to the historical values, and (2) Reduction in the variance of the KS measure, indicating that as the group size increase it matters less which judges are aggregated.

Note that some groups of size 2 and 4 have KS measures worse than the single worst judge for the parameter aggregation. Those groups include one judge with extremely large values of (α_1, β_1) for its fitted beta distribution, compared to all other judges. Beta (α_1, β_1) fits that particular judge's data well, so the quality measures for this single judge appear normal. However, group quality measures, which combine (α_1, β_1) with other judges' distribution parameters, are distorted.

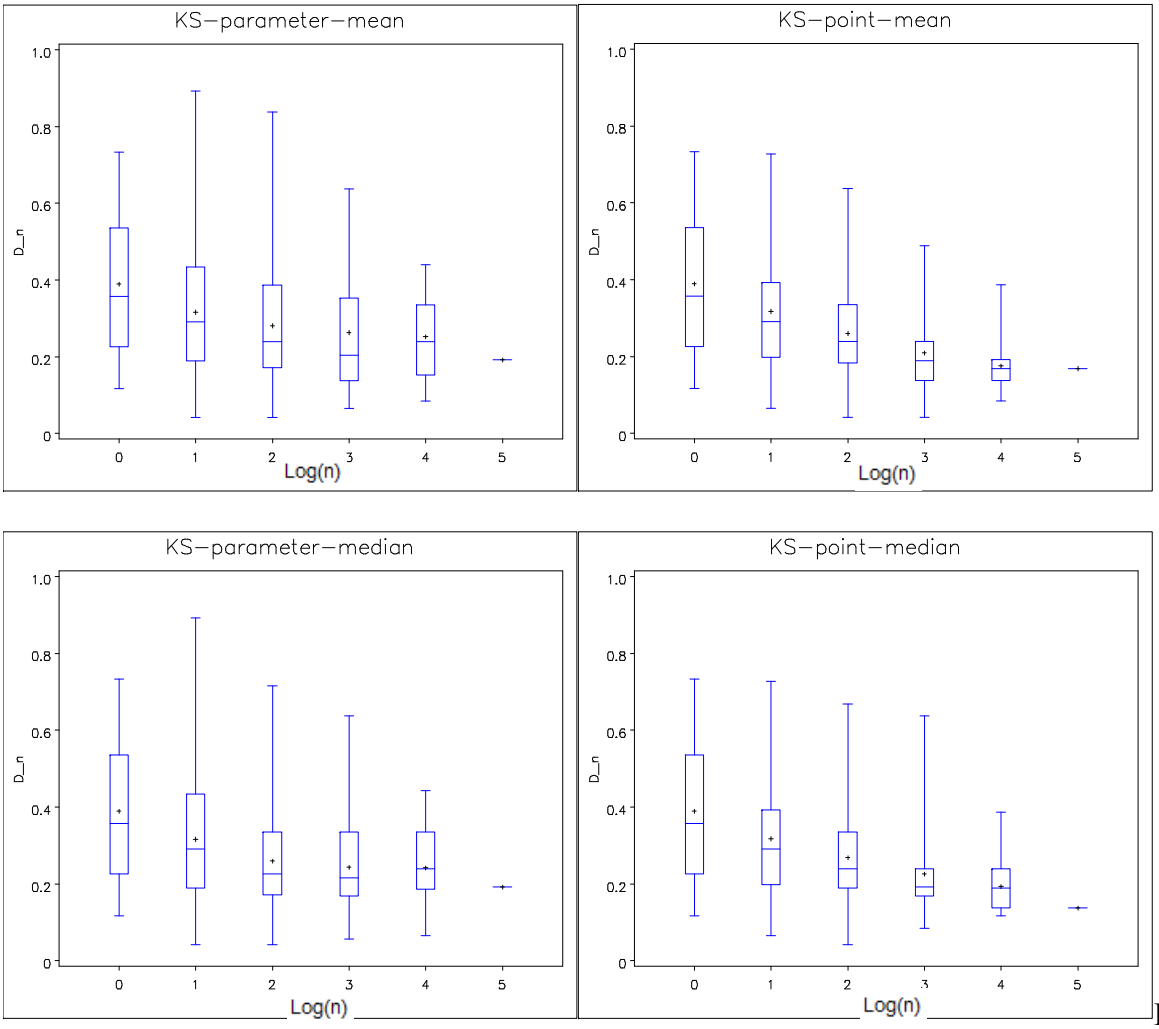


Figure 4: Distribution of KS for various aggregation methods and statistics as a function of group size for the FP assessments

Table 3 compares aggregation methods and statistics. The top two panels compare the two aggregation statistics for each method of aggregation, and the bottom two panels compare the two aggregation methods for each aggregation statistic. In each case the modal pattern is in bold face.

The top two panels show that when we aggregate points, mean aggregation is superior to median aggregation. The advantage of mean aggregation becomes larger when group size increases. When we aggregate parameters, the difference between the mean and median aggregation are relatively small, with inconsistency across groups of various sizes.

We found an interesting trend in the lower two panels. When the group size is small (≤4 judges per group) aggregating parameters results in smaller KS than aggregating points, regardless of whether we use the mean or the median,. However, the pattern is reversed when the group size increases (>4 judges per group).

Aggregating Points (%)							
Group Size	1	2	4	8	16	32	Total
KS_mean < KS_median	0	0	**42**	**45**	**46**	**100**	39
KS_mean = KS_median	**100**	**100**	24	26	35	0	**47**
KS_mean > KS_median	0	0	35	29	19	0	14

Aggregate Distribution Parameters (%)							
Group Size	1	2	4	8	16	32	Total
KS_mean < KS_median	0	0	**44**	**47**	45	0	23
KS_mean = KS_median	**100**	**100**	14	10	8	**100**	**55**
KS_mean > KS_median	0	0	42	42	**47**	0	22

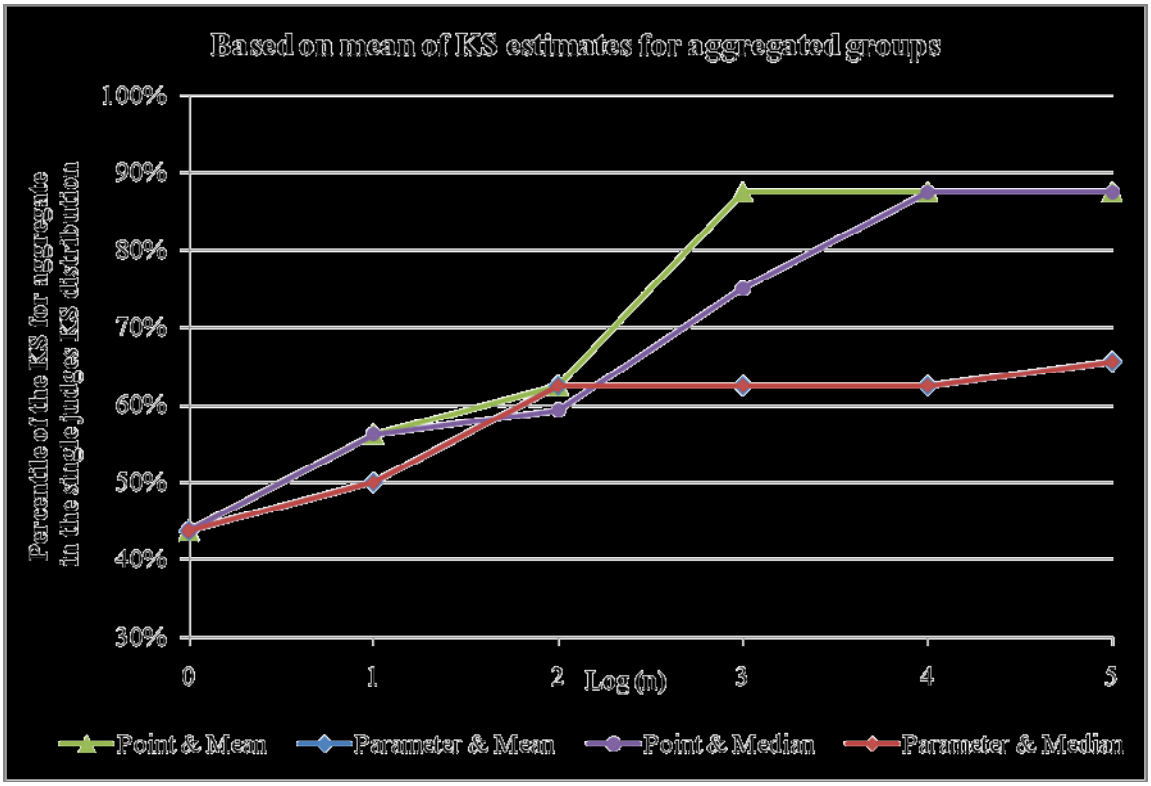
Aggregate using mean (%)							
Group Size	1	2	4	8	16	32	Total
KS_point < KS_parameter	22	38	39	**46**	**61**	**100**	**51**
KS_point = KS_parameter	34	13	14	18	19	0	16
KS_point > KS_parameter	**44**	**49**	**48**	37	20	0	33

Aggregate using median (%)							
Group Size	1	2	4	8	16	32	Total
KS_point < KS_parameter	22	38	41	**47**	**63**	**100**	**52**
KS_point = KS_parameter	34	13	10	14	16	0	14
KS_point > KS_parameter	**44**	**49**	**49**	39	22	0	34

Table 2: KS - Comparison of the two aggregation methods and the two statistics for the FP assessments (The modal pattern is in bold in each column)

Figure 5 plots the mean (top panel) and median (lower panel) of the KS distribution of the various aggregates relative the distribution of the KS of the 32 individual judges. The higher the % of single judges with KS larger than the mean KS of aggregates, the better the quality of the aggregation. In other words, the higher the line lies in Figure 5, the better quality of aggregation achieved by the corresponding aggregation method and statistic. Both panels show that mean aggregation has either lower or the same percentage value comparing to median

aggregation, and the lines for point aggregation lie above the lines for parameter aggregation, except group size of four. Overall, Figure 5 shows that mean aggregation and point aggregation leads to higher aggregation quality. We also notice that all lines level off at n=8 suggesting that there is little to gain beyond this point.



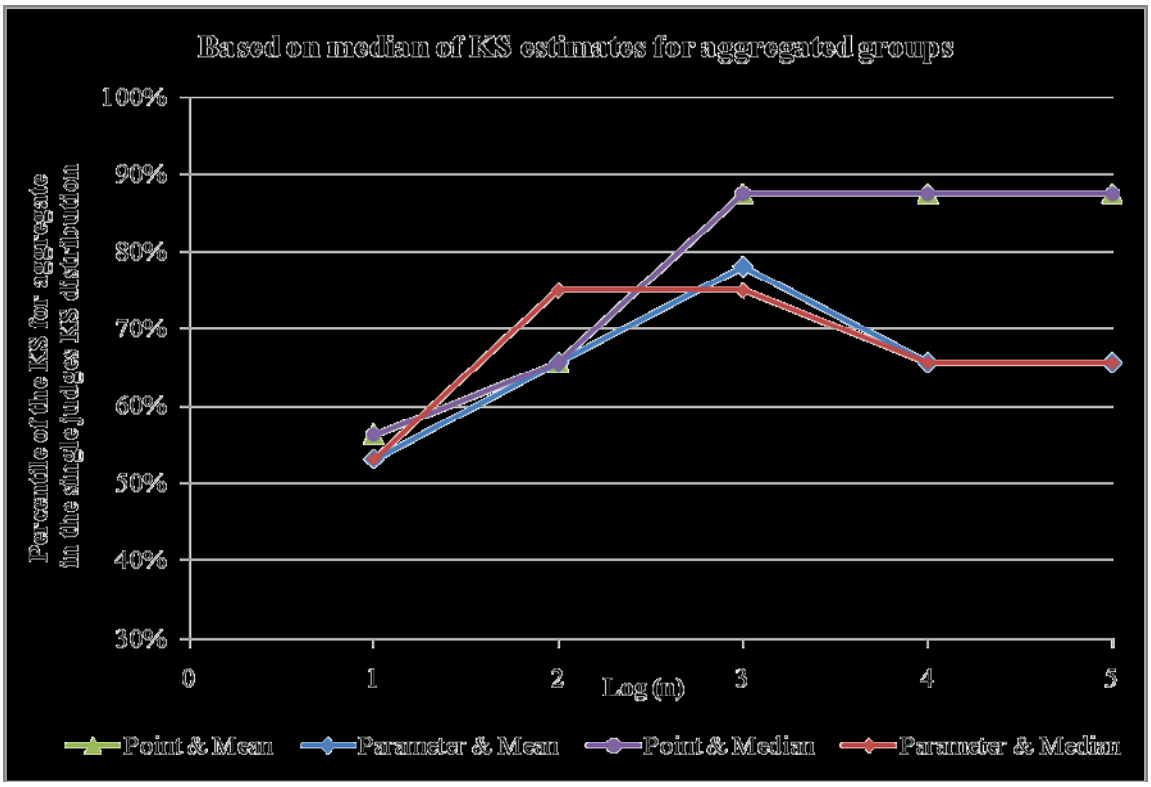


Figure 5: KS – The mean and median of the aggregated judgments compared to the distribution of single judges as a function of group size for the FP assessments

We fitted a series of nested regression models to study in more detail the major factors that affect aggregation quality. The quality measure, KS, was used as the dependent variable whereas variations of group size (n), the aggregation method, AM, (point or parameter), the aggregation statistic, AS (mean or median), and the interactions between these factors served as predictors. The standardized estimates of regression parameters and R^2 of each model are presented in Table 4.

KS (significance level: 0.05)										
MODEL NO.	PREDICTORS IN MODEL									
	Log(n)	Log^2 (n)	Method	Statistic	AM* Log(n)	AM* Log^2 (n)	AS* Log(n)	AS* Log^2 (n)	R^2	**DF**
1	-0.069	0.007							0.542	**2, 3997**
2	-0.069	0.007	0.032						0.633	**3, 3996**
3	-0.069	0.007		ns					0.542	**3, 3996**
4	-0.071	0.007	0.032	ns					0.633	**4, 3995**
5	-0.07	0.006			0.004	0.002			0.685	**4, 3995**
6	-0.067	0.006					-0.005	0.001	0.545	**4, 3995**
7	-0.069	0.005			0.004	0.002	-0.004	0.001	0.687	**6, 3993**
8	-0.083	0.007	-0.032		0.027	-0.002			0.689	**5, 3994**
9	-0.065	0.006		ns			0.004	0.002	0.545	**5, 3994**
10	-0.078	0.007	-0.032	ns	0.027	-0.002	-0.009	0.002	0.691	**8, 3991**

Table 3: Nested Regression Models for KS for the FP assessments

As group size increases, KS decreases. From the pattern of significance of the parameters' estimates (p-values) and R^2, we conclude that aggregation statistic (mean vs. median) does not have much impact on aggregation quality. Conversely, point aggregation led to a lower intercept of KS than parameter aggregation. According to this regression analysis, a higher aggregation quality is achieved by including larger number of judges in a group and aggregating points.

Similarity to the reference distribution – Measures of global quality of aggregates

We fitted beta distributions to the reference data, single judge's data, and aggregated data and compared the parameters of all these distributions.

Figure 6 shows the parameter space defined by EV and SD with a point for each of the following 53 distributions: Reference (1), each individual judge (32), the mean (or median) of all cases with n=2, 4, 8, 16, 32 for each of the four methods (point vs. parameters, mean vs. median) (20).

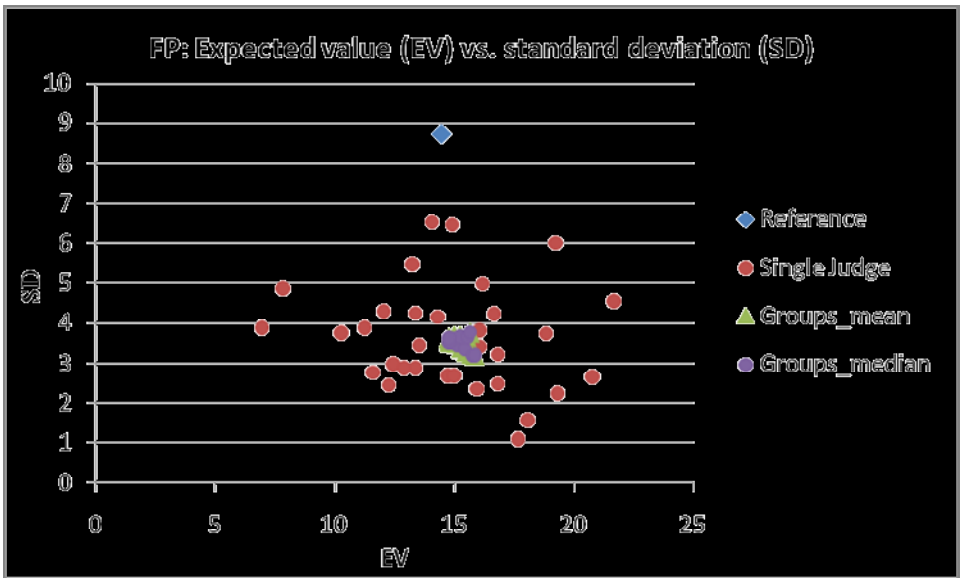


Figure 6: EV vs. SD for 53 distributions for the FP assessments

Figure 6 shows that the expected value of aggregated distributions approaches the expected value of the reference distribution. However the SDs of aggregated distributions are not as good as some of the single judge's estimates. Considering it is difficult to identify those single judges whose estimates are closer to reference data, aggregation does improve the quality of estimates.

We also plotted box plots as functions of the group size the following distributions:

- Distance (Estimated EV, Reference EV) – Figure 7
- Distance (Estimated SD, Reference SD) – Figure 8
- Euclidean Distance (Estimated (α,β), Reference (α,β)) – Figure 9

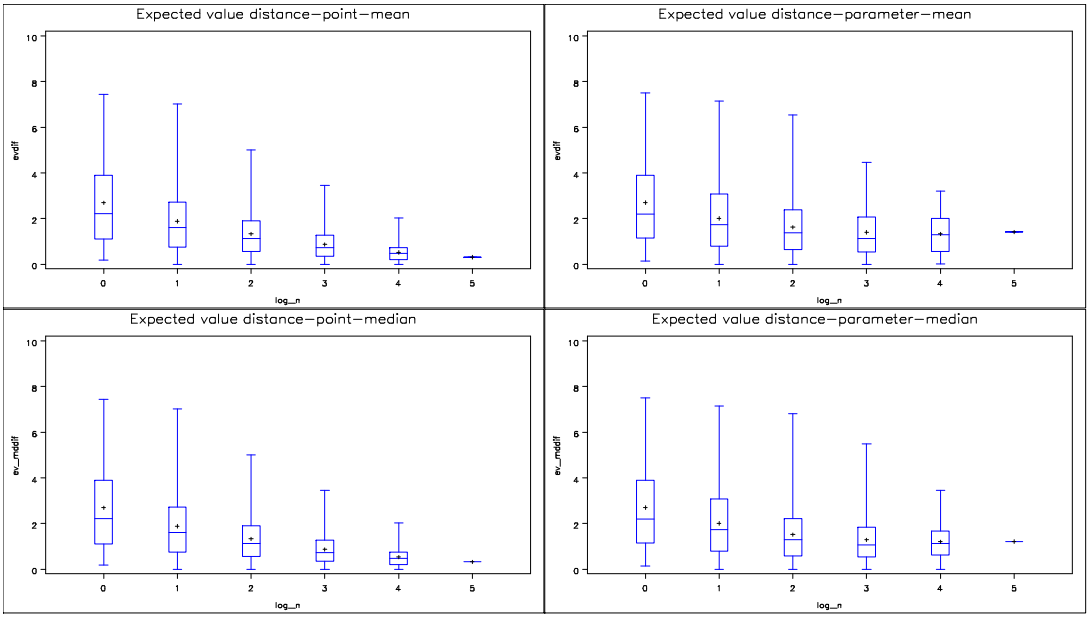


Figure 7: Distribution of distance between estimated EV and reference EV by group size for the FP assessments

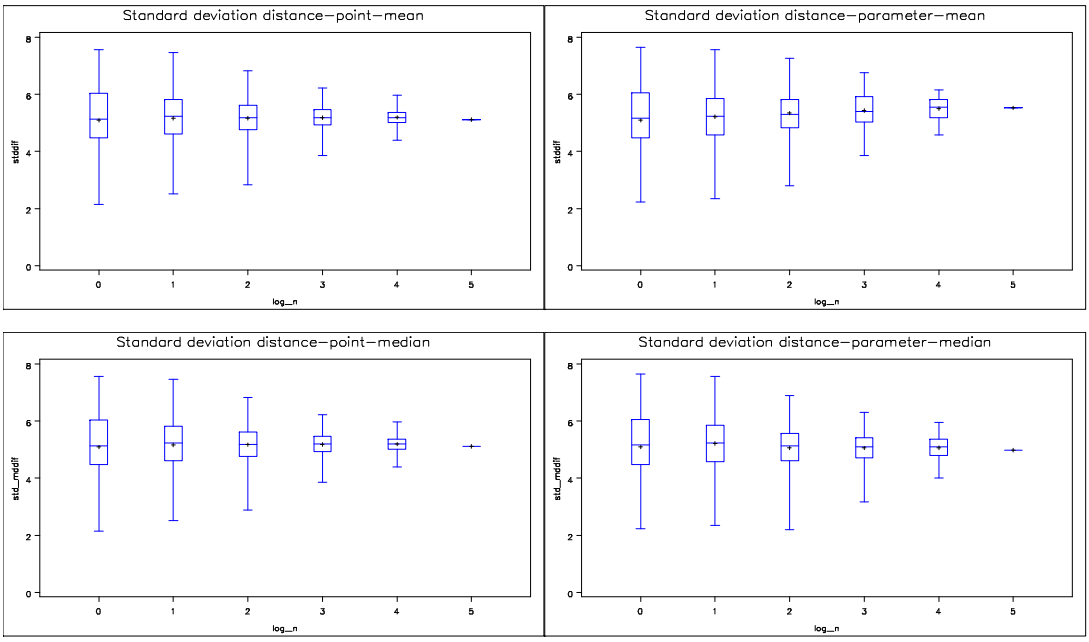


Figure 8: Distribution of distance between estimated SD and reference SD by group size for the FP assessments

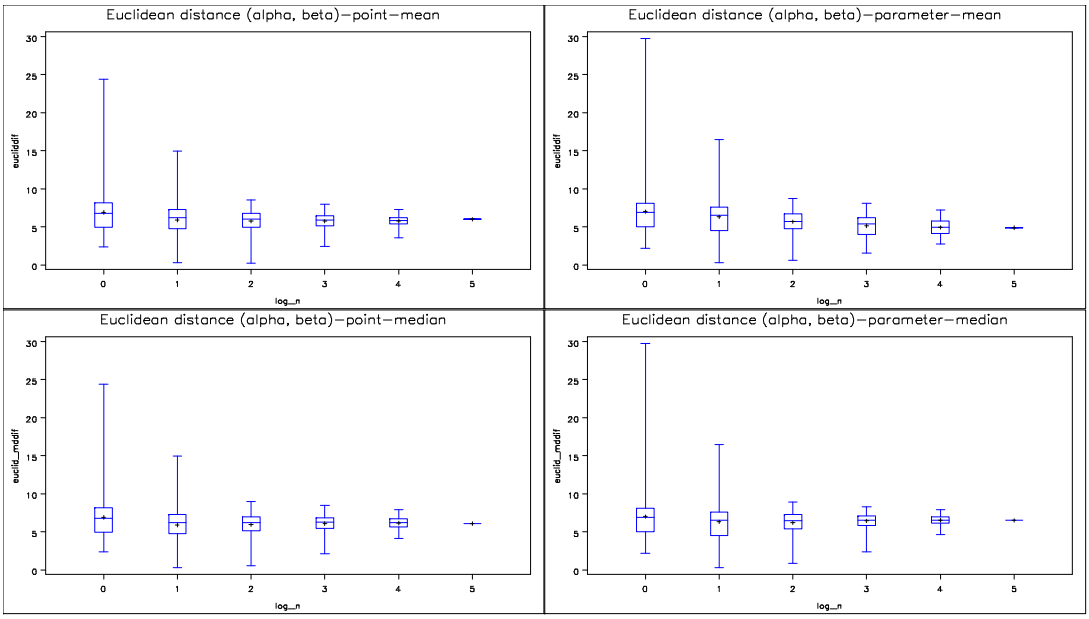


Figure 9: Distribution of Euclidian distance between estimated parameters and reference parameters by group size for the FP assessments

In Figure 7 both mean value and variance of distances for expected values decrease as group size increases. The decrease in distance is larger for point aggregation. In Figure 8 and Figure 9, there is a consistent pattern for SD distance and Euclidian distance of parameters across the four methods (point vs. parameters, mean vs. median). As group size increases, the variance of distances between estimated parameters and reference parameters becomes smaller; however the mean of distances remain about the same.

Fixed Variable Analysis

Similar analyses were performed for fixed variable approach as well.

The effects of aggregation on the range of the distributions

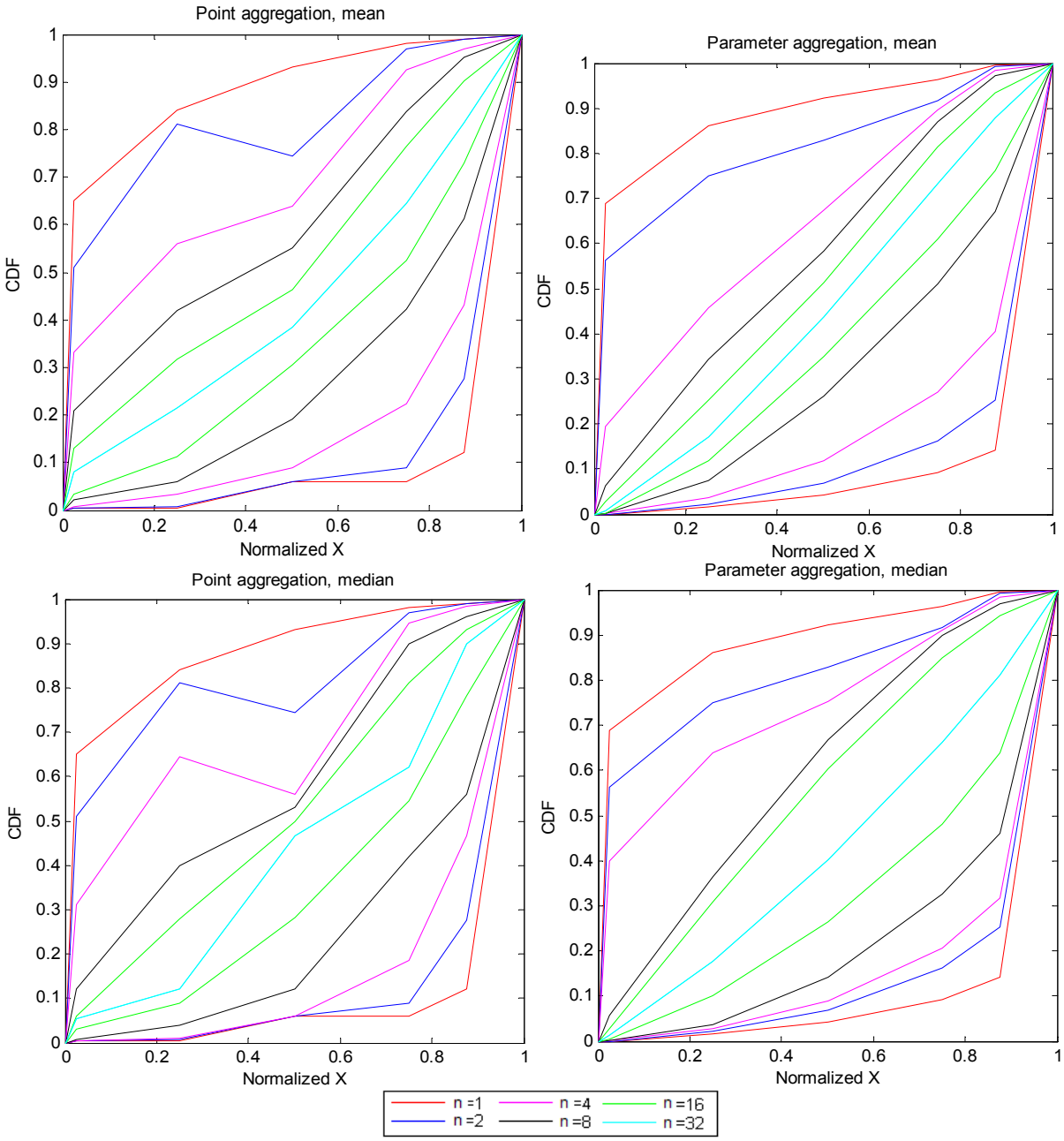


Figure 10: Bounding CDF distributions constructed from maximum and minimum CDF values of aggregated distributions as a function of group size for the FV assessments

In Figure 10 we display the maximal and minimal CDF value for each aggregated distribution (Like in Figure 3). The range of the aggregated distributions decreases monotonically as the number of distributions being aggregated increases, regardless of whether we used the mean or median, and point or parameter aggregation methods. However, some of the estimates given by

judges are not monotonic, resulting in the V shape in the left panel of Figure 10.

Similarity to the reference distribution – Kolmogorov-Smirnov analyses

Figure 11 shows that both the mean and variance of KS decreases as group size increases.

Table 5 shows that with point aggregation, median is a better statistic, while with parameter aggregation, the mean is superior. On the other hand a higher percentage of KSs using parameter aggregation results in lower KS with the mean and median aggregation method.

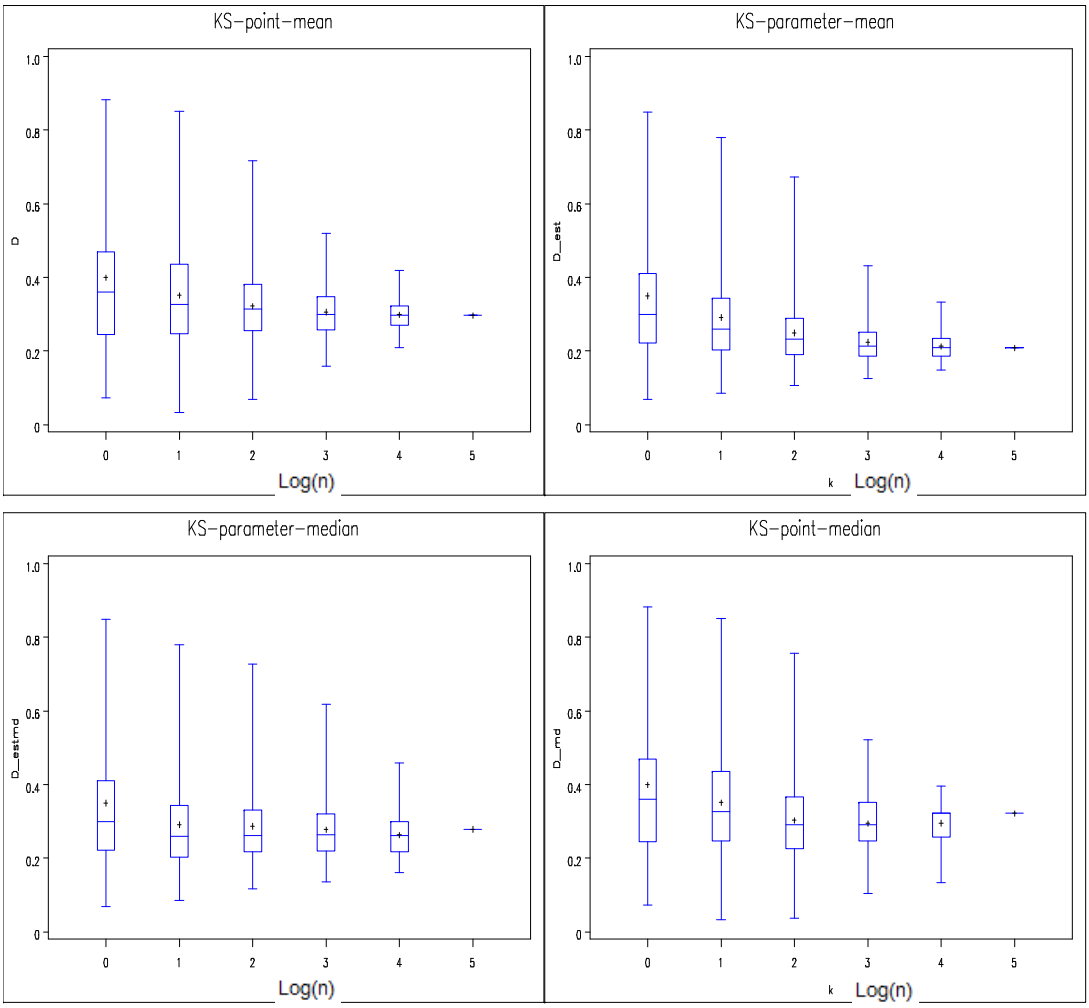


Figure 11: Distribution of KS for various aggregation methods and statistics as a function of group size for the FV assessments

Aggregating Points (%)							
Group Size	1	2	4	8	16	32	Total
KS_mean < KS_median	0	0	37	42	49	**100**	**38**
KS_mean = KS_median	**100**	**100**	2	0	1	0	34
KS_mean > KS_median	0	0	**61**	**58**	**51**	0	28

Aggregate Distribution Parameters (%)							
Group Size	1	2	4	8	16	32	Total
KS_mean < KS_median	0	0	**72**	**84**	**93**	**100**	**58**
KS_mean = KS_median	**100**	**100**	0	0	0	0	33
KS_mean > KS_median	0	0	28	16	8	0	9

Aggregate using mean (%)							
Group Size	1	2	4	8	16	32	Total
KS_point < KS_parameter	28	15	6	1	0	0	8
KS_point > KS_parameter	**72**	**85**	**94**	**99**	**100**	**100**	**92**

Aggregate using median (%)							
Group Size	1	2	4	8	16	32	Total
KS_point < KS_parameter	28	15	39	37	23	0	24
KS_point > KS_parameter	**72**	**85**	**61**	**63**	**78**	**100**	**76**

Table 4: KS - Comparison of the two aggregation methods and the two statistics for the FV assessments (The modal pattern is in bold in each column)

Figure 12 indicates that, compared to single judge, mean aggregation leads to either better or the same results as median aggregation. Parameter aggregation is in general superior.

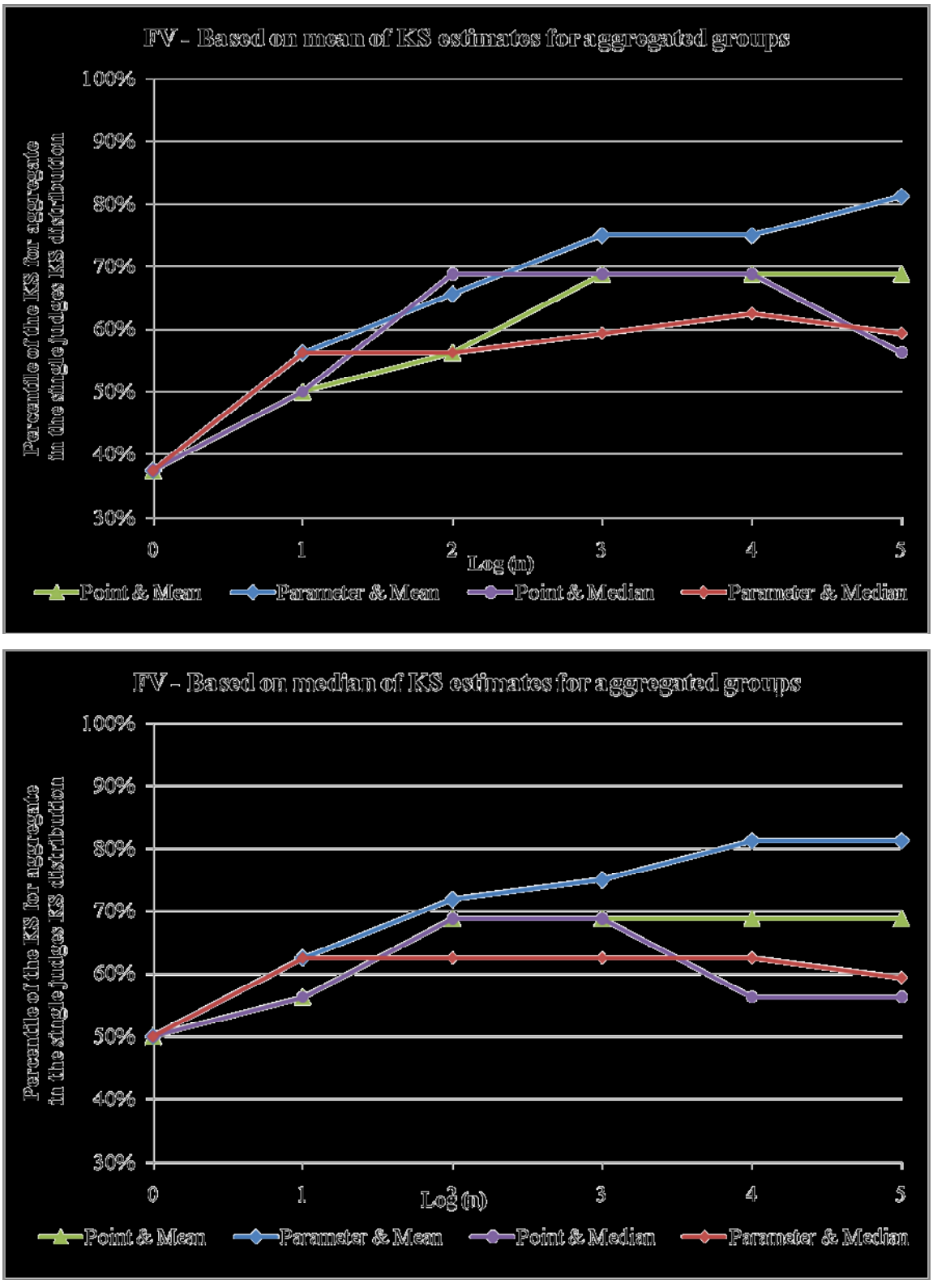


Figure 12: KS – The mean and median of the aggregated judgments compared to the distribution of single judges as a function of group size for the FV assessments

Regression analysis (Table 6) confirms the results in Figure 12 that mean aggregation yielded lower intercept of KS when the interaction of various factors are not considered. Parameter aggregation yielded lower intercept of KS as well.

	KS (significance level: 0.05)									
MODEL NO.	**PREDICTORS IN MODEL**								**R^2**	**DF**
	Log(n)	Log^2 (n)	Method	Statistic	AM* Log(n)	AM* Log^2 (n)	AS* Log(n)	AS* Log^2 (n)		
1	-0.048	0.006							0.217	**2,3997**
2	-0.048	0.006	-0.055						0.674	**3, 3996**
3	-0.048	0.006		0.020					0.279	**3, 3996**
4	-0.048	0.006	-0.055	0.020					0.736	**4, 3995**
5	-0.032	0.004			-0.033	0.004			0.650	**4, 3995**
6	-0.049	0.005					ns	0.002	0.317	**4, 3995**
7	-0.032	0.003			-0.033	0.004	ns	0.002	0.749	**6, 3993**
8	-0.057	0.008	-0.070		0.017	-0.003			0.682	**5, 3994**
9	-0.049	0.005		ns			ns	0.001	0.317	**5, 3994**
10	-0.058	0.007	-0.070	ns	0.017	-0.003	-0.009	0.001	0.781	**8, 3991**

Table 5: Nested Regression Models for KS for the FV assessments

Similarity to the reference distribution – Measure of global quality of aggregates

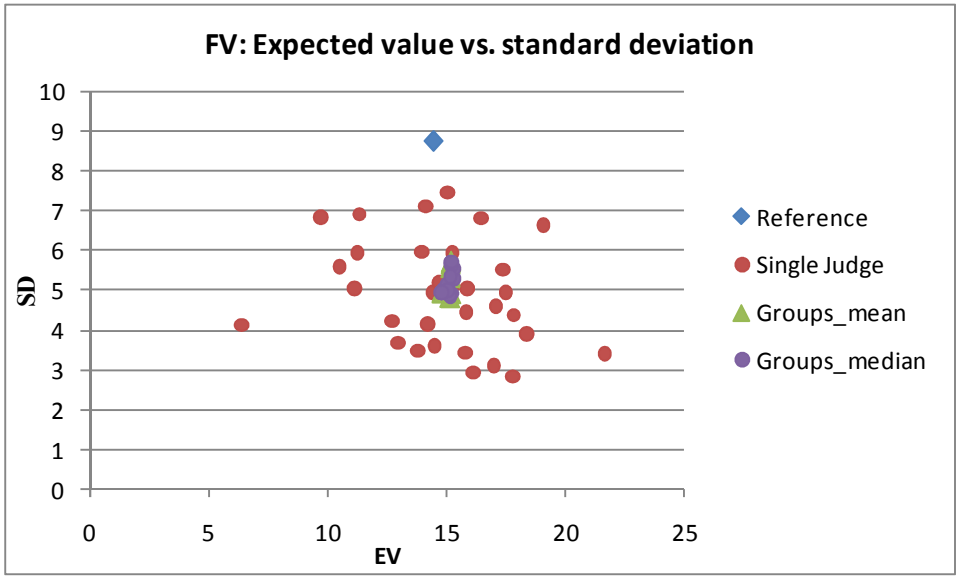


Figure 13: EV vs. SD for 53 Distributions for FV assessments

Figure 13 shows the that the expected value of aggregated distributions approaches the expected

value of the reference distribution. However the SDs of aggregated distributions are not as good as some of the single judge's estimates. This is the same pattern observed for FP in Figure 6.

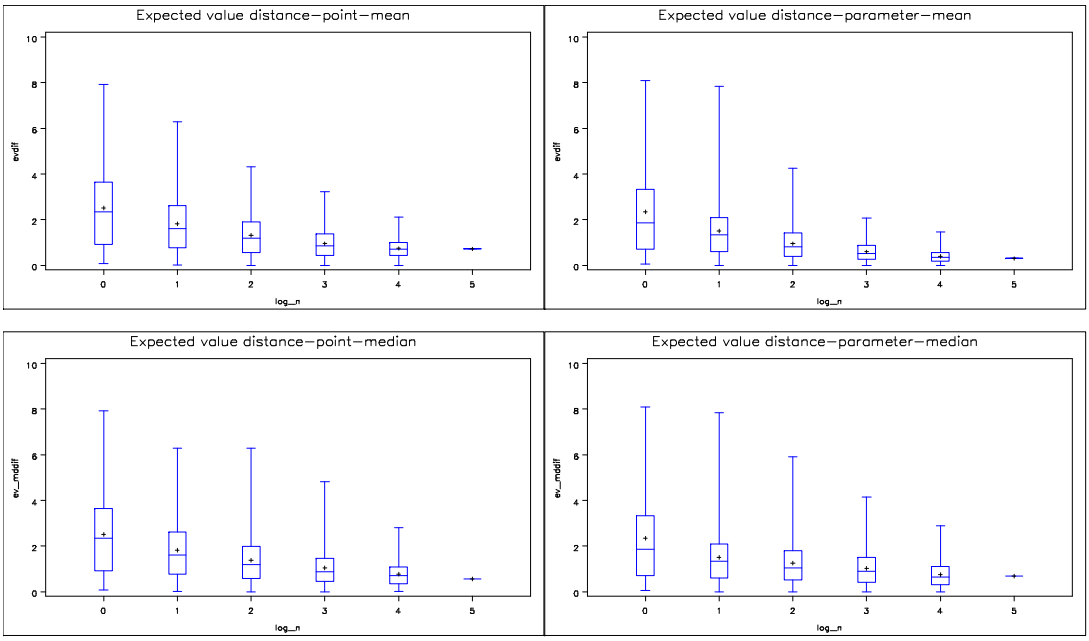


Figure 14: Distribution of distance between estimated EV and reference EV by group size for the FV assessments

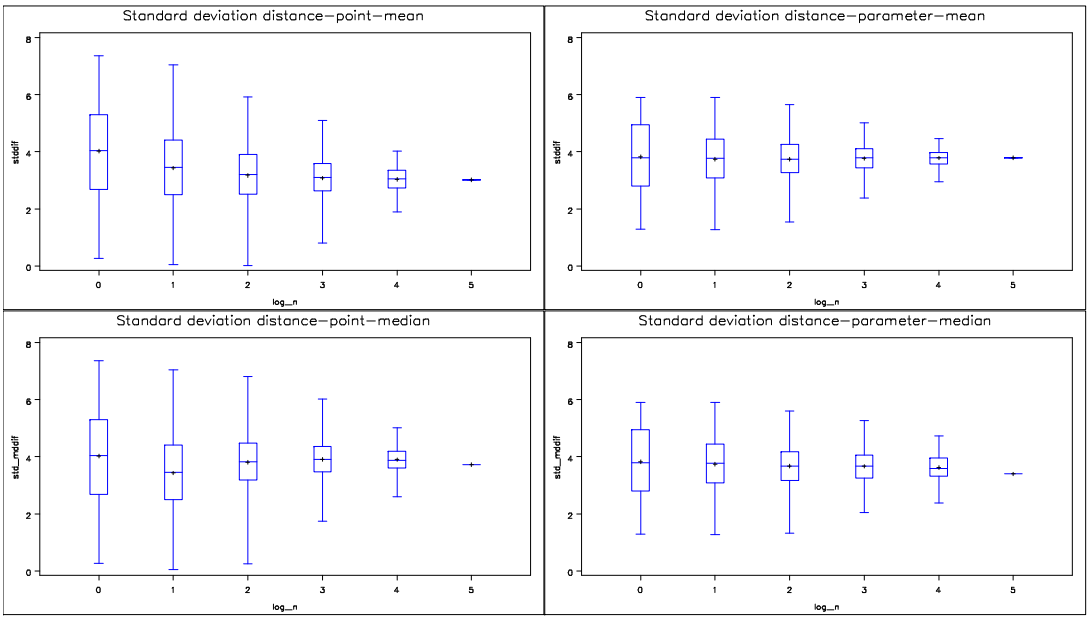


Figure 15: Distribution of distance between estimated SD and reference SD by group size for the FV assessments

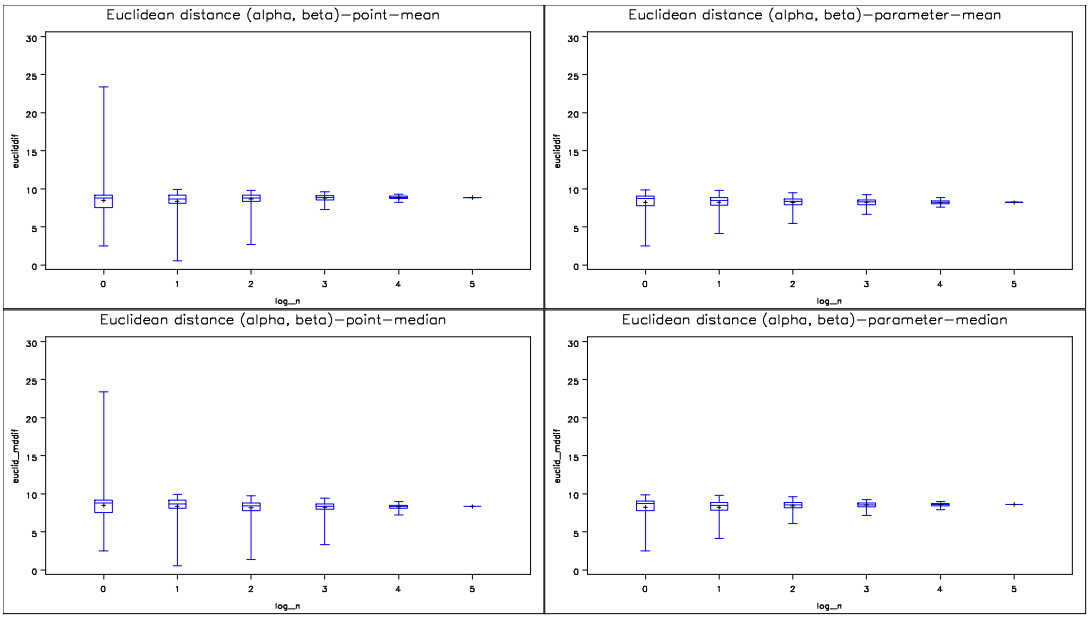


Figure 16: Distribution of Euclid distance between estimated parameters and reference parameters by group size for the FV assessments

Figures 14, 15, and 16 show results highly similar to those reported for FP (see Figures 7, 8 and 9, respectively). Both the means and the variance of distances for expected values decrease as group size increases. The decrease in distance is about the same for point and parameter aggregation, which is different from FP results. As group size increases, the variance of distances between estimated parameters (estimated SD) and reference parameters (reference SD) become smaller; however the mean of distances remain about the same.

FP vs. FV

In this section we compare directly the FP and the FV assessments. We plotted Figure 17, 18, and 19 using data from previous figures to provide a straightforward visual comparison of the two assessments. All measures based on FP data are plotted in various shades of blue, while measures based on FV data are plotted in red.

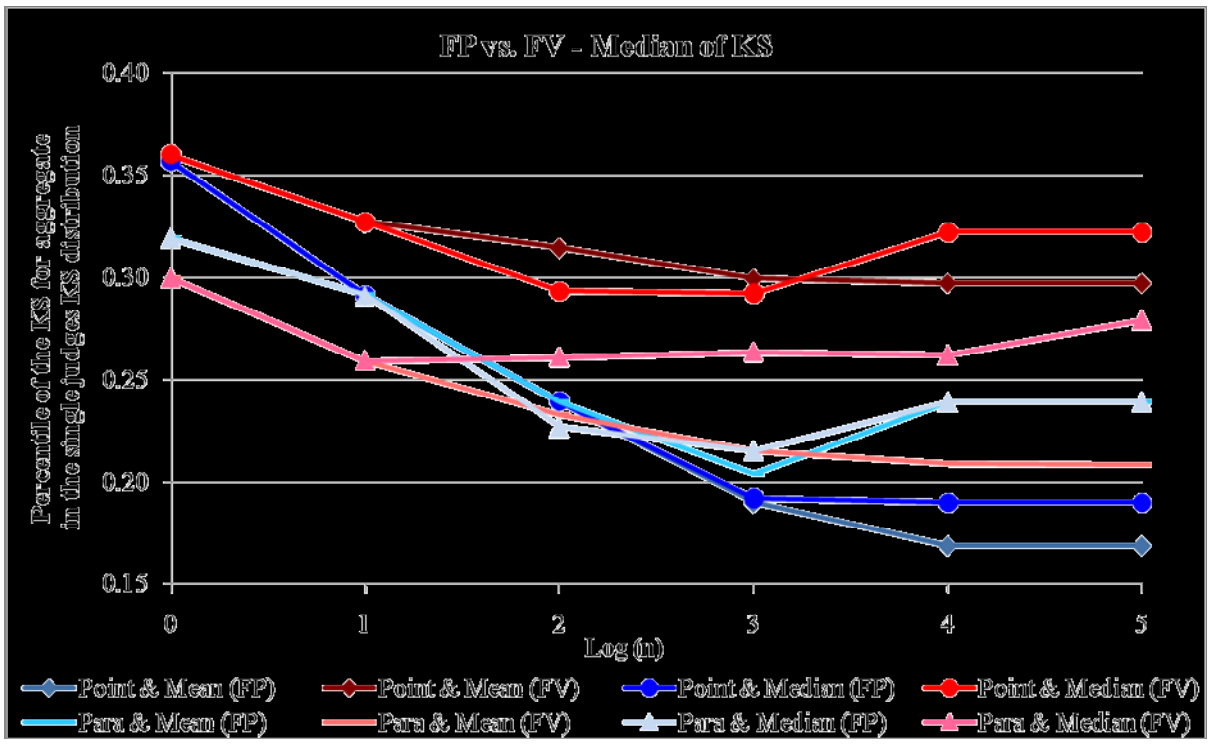


Figure 17 Median of KS for various aggregation methods and statistics as a function of group size (Compare FP and FV assessments)

Figure 17 takes the median of KS values from Figures 4 and 11. Under both assessments, KS decreases as group size increases, with a few exceptions. Exceptions exist when:

Exceptions			
FP	Parameter	Mean	Group size = 16 and 32
	Parameter	Median	Group size = 16 and 32
FV	Point	Median	Group size = 16 and 32
	Parameter	Median	Group size =16

Table 6 Exceptions Summary for Figure 17 (Compare FP and FV assessments)

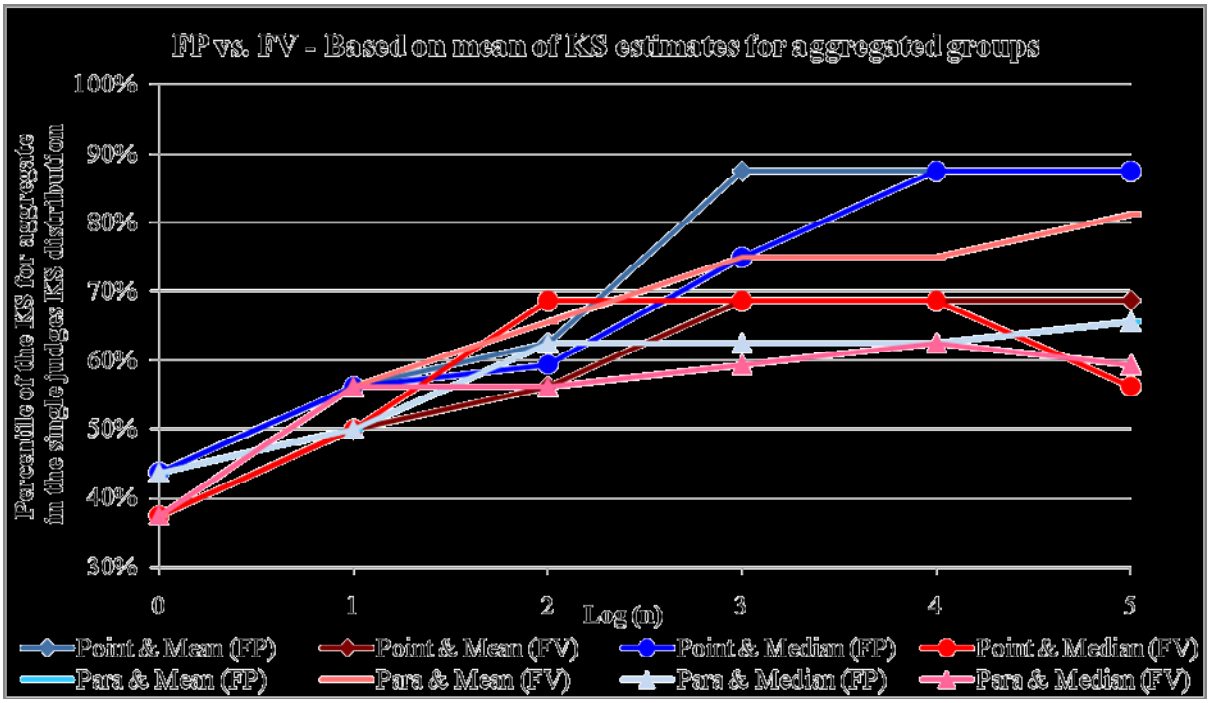
In general, results based on FP assessments outperform their FV counterparts under various combinations of aggregation methods and statistics (where blue lines lie below red lines, except for parameter aggregation using mean (where red line lies below blue line)

Within FP assessments, for all group sizes, point aggregation using mean generates the best (lowest) KS, while point aggregation using median is the second best choice. Parameter

aggregation using mean or median provide lower aggregation quality as measured by KS.

Within FV assessments, for all group sizes, parameter aggregation using mean generates the best (lowest) KS, while parameter aggregation using is the second best choice. Point aggregation using mean or median provide lower aggregation quality as measured by KS.

Considering the exception cases and the trend within each assessment, we find that point aggregation works better for FP, while parameter aggregation works better for FV.



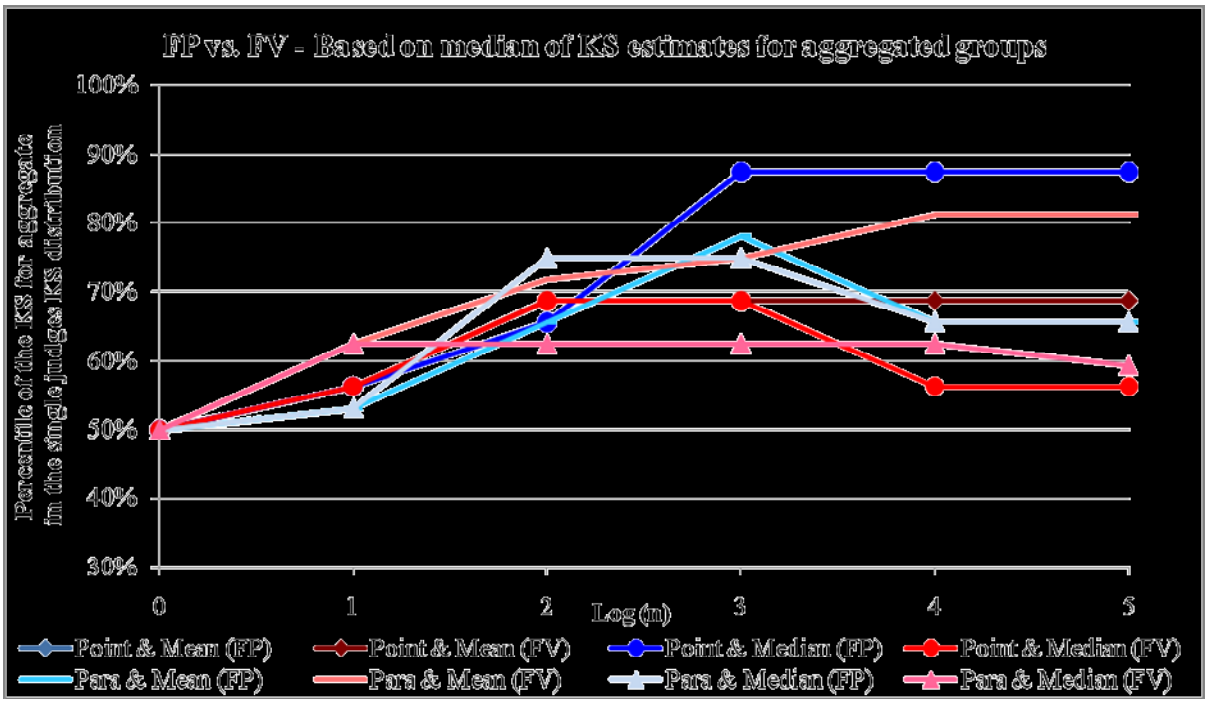


Figure 18: KS – The mean and median of the aggregated judgments compared to the distribution of single judges as a function of group size (Compare FP and FV assessments)

Figure 18 combines the data from Figures 5 and 12. The first panel is based on mean value of the KS while the second panel is based on median value. The lines started around 38% to 43% for the first panel, and 50% for the second panel (as expected as it is based on median value). A lower than 50% starting point for the first panel shows that less than half of the single judges have KS larger than the average of all single judges, which means the data is skewed and taking the average of all judges doesn't achieve as good aggregation quality as if median is used.

In the first panel, overall lines based on FP go up or stay flat as group size increases, while lines drop for FV data with group size equals 32 when median aggregation is used. In the second panel, under FP, lines go up or stay flat for point aggregation, and drop at group size equals 16 for parameter aggregation. When aggregating FV assessments, lines go up or stay flat for mean aggregation, while drop for median aggregation when group sizes equal 16 or 32. These results show that point aggregations serves better than parameter aggregation for FP, while median aggregation is less promising for FV than mean aggregation.

Table 7 summarizes the comparison between FP and FV assessments for each combination of

aggregation method and statistics.

	Mean KS (Top panel in Figure 18)	Median KS (Bottom panel in Figure 18)
Point & Mean	FP outperforms FV, and the advantage of FP increases as group size increases (from ~10% difference for n < 8 and 20% difference for n ≥ 8	FP outperforms FV except for n=4. FV advantage at n=4 (~2%) is very small comparing to FP advantage when n ≥ 8 (~20%)
Point & Median	FP outperforms FV except when group size equals 4. The advantage increase as group size increases (~20% when n=16, and ~30% when n=32)	FP outperforms FV except for n=4. FV advantage at n=4 (~2%) is very small comparing to FP advantage when n ≥ 8 (~20% to 30%)
Parameter & Mean	FV outperforms FP. Although FP generated better result for single judges, aggregation led to high quality for FV starting from n=2. The advantage of FV increases to ~15% when n ≥ 8	FV outperforms FP except for n=16. FP advantage at n=16 (~2%) is very small comparing to FV advantage when n ≠ 16 (~20% to 30%)
Parameter & Median	FP outperforms FV except when n=2. The advantage is relevant small comparing to other combination of aggregation methods and statistics (~7% across groups sizes)	FP outperforms FV except for n=2. FV advantage at n=2 is around 10% while the biggest FP advantages at n=4 and 16 are around 15%. Under both FP and FV, the lines drop when group size increases

Table 7 Comparison between FP and FV based on Figure 18

According to Table 7, parameter aggregation using mean is the only case where FV outperforms FP. This is consistent with results in Figure 17. The advantage of FP is larger with point aggregation (~20% to 30% than parameter aggregation using median (~7% or 15%)

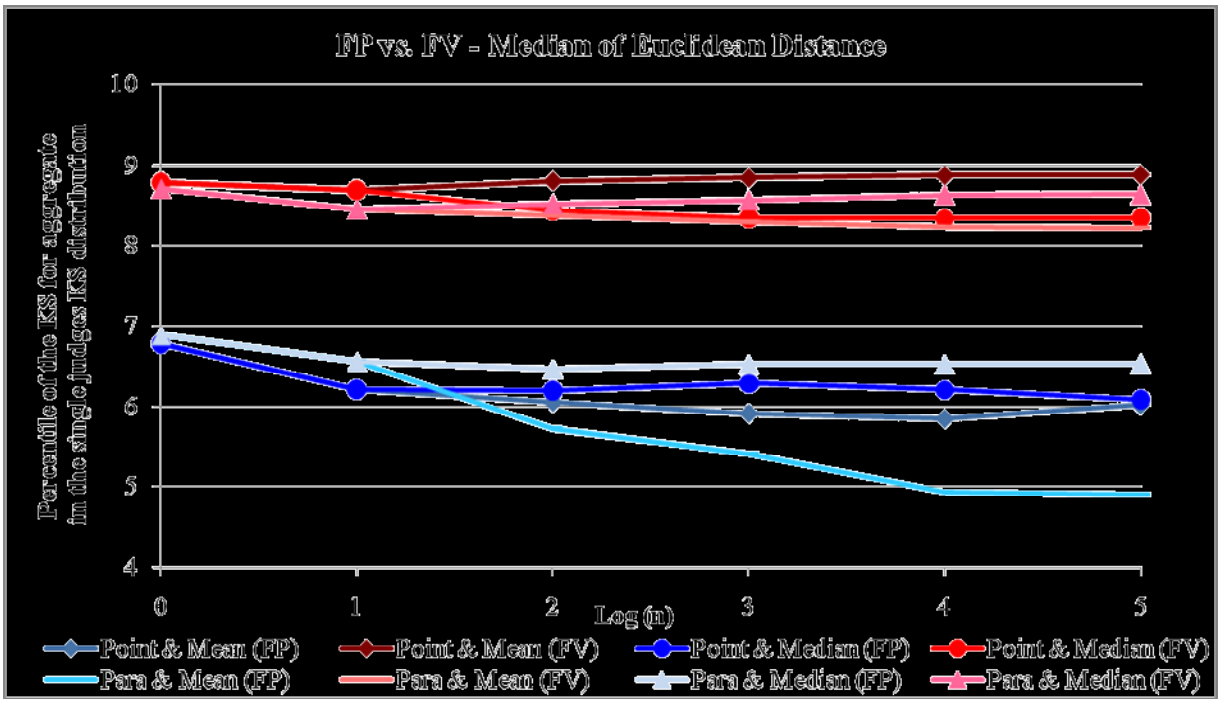


Figure 19 Median of Euclidean distance for various aggregation methods and statistics as a function of group size (Compare FP and FV assessments)

Figure 19 combines the data from Figures 9 and 16. Overall the Euclidean distances are about the same across various group sizes, with parameter aggregation using mean under FP assessment being the only exception. For all aggregation methods and statistics, Euclidean distances using FP assessments are lower than those using FV assessments, across various group sizes (the blue lines are consistently below the red lines).

Within FP or FV assessments, differences of Euclidean distances exist among various aggregation methods and statistics. However given the relevantly large variations as shown in Figures 9 and 16, we consider the difference among median values of Euclidean distances not compelling enough to be used for comparisons between aggregation methods and statistics.

To conclude, FP assessment generates higher aggregation quality under most circumstances. If FV has to be adopted for practical reasons, using parameter aggregation with mean may produce high aggregation quality. When FP is adopted, point aggregation generates better aggregation quality than parameter aggregation.

CHAPTER 4 SUMMARY AND DISCUSSION

Below is a summary of the results presented in the previous sections organized by assessment type, and quality measure.

FP assessments

1. KS:
    a. When group size increases, both the mean and the variance of KS decreases.
    b. Mean aggregation achieves results better than, or equal to, median aggregation. As group size increases, the advantage of mean aggregation increases. However the advantage is small and regression analysis shows that it is not significant.
    c. Point aggregation overall yields higher quality results than parameter aggregation, with the exception of small groups ($n \leq 4$).
2. Measure of global quality of aggregates:
    a. Aggregation decreases the distance between expected values of the estimated distribution and the reference distribution. The decrease of this distance is larger for point aggregation.
    b. As group size increases, the variance of distances between estimated parameters / estimated SD and reference parameters / reference SD become smaller. However the mean of distances remain about the same. No significant difference between aggregation methods and aggregation statistics are observed.

FV assessments

1. KS:
    a. When group size increases, both mean and variance of KS decreases.
    b. Overall mean aggregation achieves better results than median aggregation. However the advantage is small and regression analysis shows it is not significant.
    c. Parameter aggregation in general led to higher quality than point aggregation, with a few exceptions presented in Table 9.

2. Measure of global quality of aggregates:

a. Aggregation decreases the distance between expected values of the estimated distribution and the reference distribution.
b. As group size increases, the variance of distances between estimated parameters / estimated SD and reference parameters / reference SD become smaller. However the mean of distances remain about the same. No significant difference between aggregation methods and aggregation statistics is observed.

Table 8 summarizes the results above and compares the fixed probability versus fixed variable assessments.

	Fixed Probability Assessments	Fixed Variable Assessments
Group Size	As group size increases aggregation quality improves	As group size increases aggregation quality improves
Aggregation Statistics (Mean vs. Median)	With a few exceptions mean aggregation is superior to median aggregation. As group size increases, the advantage becomes larger. The superiority is not significant.	With a few exceptions mean aggregation is superior to median aggregation. The superiority is not significant.
Aggregation Method (Point vs. Parameter)	With a few exceptions, point aggregation achieves higher aggregation quality then parameter aggregation	With a few exceptions parameter aggregation shows advantage over point aggregation.

Table 8: Comparison of FP with FV using results from multiple measures

As stated in the "FP vs. FV" section, FP assessment generates higher aggregation quality than FV assessment under most circumstances. If FV has to be adopted for practical reasons, using parameter aggregation with mean may produce high aggregation quality. When FP is adopted, point aggregation generates better aggregation quality than parameter aggregation.

REFERENCES

Abbas, A.E. (2009). A Kullback-Leibler view of linear and log-linear pools. *Decision Analysis*, *6*, 25–37.

Abbas, A. E., Budescu, D.V., & Yu, H.T., & Haggerty, R. (2008). A comparison of two probability encoding methods: fixed probability vs. fixed variable values. *Decision Analysis,* 5, 190-202.

Agnew, C. (1985). Bayesian consensus forecasts of macroeconomic variables. *Journal of Forecasting*, *4*, 363−376.

Ariely, D., Au, W. T., Bender, R. H., Budescu, D. V., Dietz, C. B., Gu, H. Wallsten, T. S., & Zauberman, G. (2000). The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied*, *6*, 130–147.

Bordley R.F. (2009) Combining the opinions of experts who partition events differently. *Decision Analysis,6*, 38-46.

Broomell, S.B., & Budescu, D.V. (2009). Why are experts correlated? Decomposing correlations between judges. *Psychometrika,* 74, 531-553.

Budescu, D.V. (2006). Confidence in aggregation of opinions from multiple sources. In K. Fiedler & P. Juslin (Eds.), *Information sampling and adaptive cognition* (pp. 327–354). Cambridge: Cambridge University Press.

Budescu, D.V., & Yu, H.T. (2006). To Bayes or Not to Bayes? A comparison of two classes of models of information aggregation. *Decision Analysis*, *3*, 145-162.

Clemen, R.T., & Ulu, C. (2008). Interior additivity and subjective probability assessment of continuous variables. *Management Science*, *54*, 835-851

Clemen, R.T., & Winkler, R.L. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis*, *19*, 187-203

Clemen, R.T., & Winkler, R.L. (2007). Aggregating probability distributions. In W. Edwards, R.F. Miles, and D.von Winterfeldt, eds., *Advances in decision analysis: from foundations to applications* (pp. 154–176). Cambridge: Cambridge University Press.

Cooke, R.M. (1991). *Experts in uncertainty: Opinion and subjective probability in science.* New York: Oxford University Press.

Engelberg, J., Manskiy, C. F., & Williams, J. (2007). Comparing the point predictions and subjective probability distributions of professional forecasters. *National Bureau of Economic Research Inc Working Papers*, *11978.*

Fiedler, K. (1996). Explaining and simulating judgment biases as an aggregation phenomenon in probabilistic multiple-cue environments. *Psychological Review*, *103*, 193–214.

French, S. and Insua, D. (2000). *Statistical decision theory.* London: Arnold.

Genest, C., & Zidek, J.V. (1986). Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, *1*, 114-135.

Guerard Jr.J.B. & Clemen, R.T. (1989). Collinearity and the use of latent root regression for combining GNP forecasts. (Gross National Product). *Journal of Forecasting*, *8*, 231-238

Harvey, N., Bolger, F., & McClelland, A. (1994). On the nature of expectations. *British Journal of Psychology*, *85*, 203 -230.

Heathcote, A., Popiel, S. J., & Mewhort, D. J. (1991). Analysis of response time distributions: An example using the Stroop task. *Psychological Bulletin*, *109*, 340-347.

Hora, S. C., Hora, J. A., & Dodd, N. G. (1992). Assessment of probability distribution for continuous random variables: a comparison of bisection and fixed value methods. *Organizational Behavioral Human Decision Process*, *51*, 135-155.

Jiang, Y., Rouder, J. N., & Speckman, P. L. (2004). A note on the sampling properties of the Vincentizing (quantile averaging) procedure. *Journal of Mathematical Psychology*, *48*, 186-195

Johnson, T.R., Budescu, D.V., & Wallsten, T.S. (2001). Averaging probability judgments: Monte Carlo analyses of asymptotic diagnostic value. *Journal of Behavioral Decision Making*, *14*, 123–140.

Jose, V.R.R., & Winkler, R.L. (2008). Simple robust averages of forecasts: Some empirical results. *International Journal of Forecasting*, *24*, 163–169

Lau, A. H. L., Lau, H. S., & Zhang, Y. (1996). A simple and logical alternative for making PERT time estimates. *IIE Transactions, 28*, 183-192.

Lehmann, E. L. (2004). *Elements of Large Sample Theory* (pp. 343). Springer-Verlag New York, LLC.

Lichtenstein, S., Fichhoff, B., & Phillips, L. D. (1982) Calibration of probabilities: the state of the art to 1980. In D. Kahneman, P. Slovic, and A. Tversky, eds., *Judgment under Uncertainty: Heuristics and Biases* (pp. 306-334). Cambridge: Cambridge University Press.

McNamee, P., & Celona, J. (2001). *Decision analysis for the professional*, 3rd ed. Menlo Park, CA: SmartOrg Inc.

McNees, S. K. (1992). The uses and abuses of 'consensus' forecasts. *Journal of Forecasting*, *11*, 703−711.

Morris, P.A. (1986). Comment on Genest and Zideck's "Combining probability distributions: A critique and annotated bibliography". *Statistical Science*, *1*, 141–144.

Rouder, J. N., Lu, J., Speckman, P., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review*, *12,* 195-223.

Spetzler, C. S., & Stael Von Holstein, C.-A. S. (1975). Probability encoding in decision analysis. *Management Science*, *22*, 340-358.

Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, *69 (347)*, 730- 737.

Winkler, R.L. (1981). Combining probability distributions from dependent information sources. *Management Science*, *27*, 479–488.

Stock, J. H., & Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, *23*, 405−430.

Stone, M. (1961). The opinion pool. *Annals* of *Mathematical Statistics*, *32*, 1339-1342.

Thomas, E.A.C., & Ross, B.H. (1980). On Appropriate procedures for combining probability distributions within the same family. *Journal of Mathematical Psychology*, *21*, 136-152

Vincent, S. B. (1912). The function of vibrissae in the behavior of the white rat. *Behavioral Monographs*, *1 (5).*

Wallsten, T. S., Diederich, A. (2001). Understanding pooled subjective probability estimates. *Mathematical Social Sciences*, *41*, 1–18.

Wallsten, T.S., Forsyth, B. & Budescu, D.V. (1983). Stability and coherence of health experts' upper and lower subjective probabilities about dose-response curves. *Organizational Behavior and Human Decision Processes*, 31, 277-302.

Yaniv, I.,&Kleinberger, E. (2000). Advice taking in decision-making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, *83*, 260–281.

APPENDIX

Table 9 shows the summary of cases which were compared in this study.

				Points		Parameters	
Group Size	No. of Groups	Repetitions	Total cases	Mean	Median	Mean	Median
1	32	1	1				
2	16	200	3200				
4	8	200	1600				
8	4	200	800				
16	2	200	400				
32	1	1	1				
Ref	1	1	1				
All	64		6003				

Table 9: Summary of cases compared

Table 10 and 11 show the experiment data for two judges (judge 1 and judge2), the aggregated data for the group of these two judges (group1), and the calculation of the quality measurements.

For single judges (judge 1 and 2 in this case) – FP and FV

1) max, min: upper and lower bound of the temperature given by judges at the beginning of the experiment.
2) a, b: alpha and beta estimates for fitted beta distribution to each judge. For FP, beta distribution is fitted to X and F, where F is a fixed set of values given to all judges; For FV, beta distribution is fitted to F and N(X), where N(X) is a fixed set of values across judges
3) a_ref, b_ref: alpha and beta parameters of the reference distribution
4) X_est: estimated X on the fitted beta distribution corresponding to a given F value (FP)
5) F_est: estimated F on the fitted beta distribution corresponding to a given X value (FV)
6) KS(point): KS measures based on X and F values (FP) or N(X) and F values (FV)
7) KS(parameter): KS measures based on X_est and F values (FP) or N(X) and F_est values (FV)
8) EV: expected value of the fitted beta distribution
9) SD: standard deviation of the fitted beta distribution
10) D(EV): distance between EV of estimated distribution and reference distribution
11) D(SD): distance between SD of estimated distribution and reference distribution
12) D(a,b): Euclid distance between (a,b) of estimated distribution and reference distribution
13) KS(parameter): KS measures based on X_est and F values (FP) or N(X) and F_est values (FV)

For groups (judge 1 and 2 together formed group1) - FP

1) Point / Mean: calculate KS based on F and X_mean, which is the average of X for judge 1 and judge 2; derive a_fap and b_fap by fitting beta distribution to the aggregated points (F and X_mean) ; calculate EV, SD, and distance measures based on beta(a_fap, b_fap) and beta(a_ref, b_ref).

2) Point / Median: calculate KS based on F and X_median, which is the median of X for judge 1 and judge 2; derive a_fapmd and b_fapmd by fitting beta distribution to the aggregated points (F and X_median); calculate EV, SD, and distance measures based on beta(a_fapmd, b_fapmd) and beta(a_ref, b_ref).
3) Parameter / Mean: derive a_mean and b_mean as the averaged alpha and beta of judge 1 and judge 2; infer X_est based on beta (a_mean, b_mean) with given F values; calculate KS based on F and X_est; calculate EV, SD, and distance measures based on beta(a_mean, b_mean) and beta(a_ref, b_ref)
4) Parameter / Median: derive a_median and b_median as the median of judge 1 and judge 2's alphas and betas; infer X_estmd based beta (a_median, b_median) with given F values; calculate KS based on F and X_estmd; calculate EV, SD, and distance measures based on beta(a_median, b_median) and beta(a_ref, b_ref)

For groups (judge 1 and 2 together formed group1) - FV

5) Point / Mean: calculate KS based on N(X) and F_mean, which is the average of F for judge 1 and judge 2; derive a_fap and b_fap by fitting beta distribution to the aggregated points (N(X) and F_mean); calculate EV, SD, and distance measures based on beta(a_fap, b_fap) and beta(a_ref, b_ref).
6) Point / Median: calculate KS based on N(X) and F_median, which is the median of F for judge 1 and judge 2; derive a_fapmd and b_fapmd by fitting beta distribution to the aggregated points (N(X) and F_median); calculate EV, SD, and distance measures based on beta(a_fapmd, b_fapmd) and beta(a_ref, b_ref).
7) Parameter / Mean: derive a_mean and b_mean as the averaged alpha and beta of judge 1 and judge 2; infer F_est based on beta (a_mean, b_mean) with given N(X) values; calculate KS based on N(X) and F_est; calculate EV, SD, and distance measures based on beta(a_mean, b_mean) and beta(a_ref, b_ref)
8) Parameter / Median: derive a_median and b_median as the median of judge 1 and judge 2's alphas and betas; infer F_estmd based beta (a_median, b_median) with given N(X) values; calculate KS based on N(X) and F_estmd; calculate EV, SD, and distance measures based on beta(a_median, b_median) and beta(a_ref, b_ref)

Judge 1						
max	**min**	**a**	**b**	**F**	**X**	**X_est**
19.444	8.889	1.557	1.260	0.05	11	10
				0.25	11.5	12.5
				0.5	15	15
				0.75	17	17
				0.95	19	19
KS (Point)	**KS (Parameter)**	**EV**	**SD**	**D(EV)**	**D(SD)**	**D(a,b)**
0.137	0.137	14.724	2.686	0.274	6.054	8.066

Judge 2						
max	**min**	**a**	**b**	**F**	**X**	**X_est**
30.000	6.667	1.485	1.271	0.05	9	9
				0.25	14.5	14.5
				0.5	19.5	19.5
				0.75	24.5	24
				0.95	28	28.5
KS (Point)	**KS (Parameter)**	**EV**	**SD**	**D(EV)**	**D(SD)**	**D(a,b)**
0.477	0.477	19.241	6.002	4.791	2.738	8.114

	Point Mean									
	F	**X_mean**	**KS**	**a_fap**	**b_fap**	**EV**	**SD**	**D(EV)**	**D(SD)**	**D(a,b)**
Group 1	0.05	10	0.387	1.436	1.184	17.065	4.433	2.615	4.307	8.208
	0.25	13								
	0.5	17.5								
	0.75	21								
	0.95	23.5								

	Point Median									
	F	**X_median**	**KS**	**a_fapmd**	**b_fapmd**	**EV**	**SD**	**D(EV)**	**D(SD)**	**D(a,b)**
Group 1	0.05	10	0.387	1.436	1.184	17.065	4.433	2.615	4.307	8.208
	0.25	13								
	0.5	17.5								
	0.75	21								
	0.95	23.5								

	Parameter Mean									
	F	**a_mean**	**b_mean**	**X_est**	**KS**	**EV**	**SD**	**D(EV)**	**D(SD)**	**D(a,b)**
Group 1	0.05	1.521	1.265	9.5	0.387	17.028	4.336	2.578	4.404	8.090
	0.25			13.5						
	0.5			17						
	0.75			20.5						
	0.95			23.5						

	Parameter Median									
	F	**a_median**	**b_median**	**X_estmd**	**KS**	**EV**	**SD**	**D(EV)**	**D(SD)**	**D(a,b)**
Group 1	0.05	1.521	1.265	9.5	0.387	17.028	4.336	2.578	4.404	8.090
	0.25			13.5						
	0.5			17						
	0.75			20.5						
	0.95			23.5						

Table 10: Experiment Data for two subjects and Measurement Calculation for the FP assessments

Judge 1						
max	**min**	**a**	**b**	**F**	**X**	**X_est**
19.444	8.889	0.608	0.535	0.025	0.060	0.067
				0.250	0.250	0.285
				0.500	0.500	0.462
				0.750	0.560	0.646
				0.875	0.870	0.761
KS (Point)	**KS (Parameter)**	**EV**	**SD**	**D(EV)**	**D(SD)**	**D(a,b)**
0.382	0.296	14.501	3.598	0.051	5.142	9.258

Judge 2						
max	**min**	**a**	**b**	**F**	**X**	**X_est**
30.000	6.667	1.586	1.867	0.025	0.035	0.007
				0.250	0.220	0.225
				0.500	0.470	0.568
				0.750	0.890	0.865
				0.875	0.970	0.961
KS (Point)	**KS (Parameter)**	**EV**	**SD**	**D(EV)**	**D(SD)**	**D(a,b)**
0.188	0.252	17.383	5.510	2.933	3.230	7.661

	Point Mean									
	F	**X_mean**	**KS**	**a_fap**	**b_fap**	**EV**	**SD**	**D(EV)**	**D(SD)**	**D(a,b)**
Group 1	0.025	0.048	0.217	1.096	1.065	16.370	4.765	1.920	3.975	8.544
	0.250	0.235								
	0.500	0.485								
	0.750	0.725								
	0.875	0.920								

	Point Median									
	F	**X_median**	**KS**	**a_fapmd**	**b_fapmd**	**EV**	**SD**	**D(EV)**	**D(SD)**	**D(a,b)**
Group 1	0.025	0.048	0.217	1.096	1.065	16.370	4.765	1.920	3.975	8.544
	0.250	0.235								
	0.500	0.485								
	0.750	0.725								
	0.875	0.920								

	Parameter Mean									
	F	**a_mean**	**b_mean**	**X_est**	**KS**	**EV**	**SD**	**D(EV)**	**D(SD)**	**D(a,b)**
Group 1	0.025	1.097	1.201	0.021	0.226	15.865	4.660	1.415	4.080	8.457
	0.250			0.258						
	0.500			0.533						
	0.750			0.793						
	0.875			0.909						

	Parameter Median									
	F	**a_median**	**b_median**	**X_estmd**	**KS**	**EV**	**SD**	**D(EV)**	**D(SD)**	**D(a,b)**
Group 1	0.025	1.097	1.201	0.021	0.226	15.865	4.660	1.415	4.080	8.457
	0.250			0.258						
	0.500			0.533						
	0.750			0.793						
	0.875			0.909						

Table 11: Experiment Data for two subjects and Measurement Calculation for the FV assessments